ENCYCLOPEDIA OF

MEDICAL DEVICES AND INSTRUMENTATION

Second Edition

VOLUME 4

Hydrocephalus, Tools for Diagnosis and Treatment of – Monoclonal Antibodies

ENCYCLOPEDIA OF MEDICAL DEVICES AND INSTRUMENTATION, SECOND EDITION

Editor-in-Chief
John G. Webster
University of Wisconsin–Madison

Editorial Staff
Vice President, STM Books: **Janet Bailey**
Associate Publisher: **George J. Telecki**
Editorial Director: **Sean Pidgeon**
Director, Book Production and Manufacturing:
Camille P. Carter
Production Manager: **Shirley Thomas**
Illustration Manager: **Dean Gonzalez**
Senior Production Editor: **Kellsee Chu**
Editorial Program Coordinator: **Surlan Murrell**

ENCYCLOPEDIA OF

MEDICAL DEVICES AND INSTRUMENTATION

Second Edition

Volume 4

Hydrocephalus, Tools for Diagnosis and Treatment of – Monoclonal Antibodies

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

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warrnaties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

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-13 978-0-470-04069-0 (v. 4 : cloth)
ISBN-10 0-470-04069-6 (v. 4 : 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, *Universite´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 RenéDescartes 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

Trautman, Edwin D., *RMF Strategies, Cambridge, Massachusetts*, Cardiac Output, Thermodilution Measurement of

Treena, Livingston Arinzeh, *New Jersey Institute of Technology, Newark, New Jersey*, Polymeric Materials

Trentman, Terrence L., *Mayo Clinic Scottsdale*, Spinal Cord Stimulation

Troken, Alexander J., *University of Illinois at Chicago, Chicago, Illinois*, Cartilage and Meniscus, Properties of

Tsaftaris, Sotirios A., *Northwestern University, Evanston, Illinois*, DNA Sequence

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

Tulipan, Noel, *Vanderbilt University Medical Center, Nashville, Tennessee*, Intrauterine Surgical Techniques

Tuteja, Ashok K., *University of Utah, Salt Lake City, Utah*, Anorectal Manometry

Ty, Smith N., *University of California, San Diego, California*, Physiological Systems Modeling

Tyrer, Harry W., *University of Missouri-Columbia, Columbia, Missouri*, Cytology, Automated

Valvano, Jonathan W., *The University of Texas, Austin, Texas*, Bioheat Transfer

Van den Heuval, Frank, *Wayne State University, Detroit, Michigan*, Imaging Devices

Veit, Schnabel, *Aalborg University, Aalborg, Denmark*, Electroneurography

Velanovich, Vic, *Henry Ford Hospital, Detroit, Michigan*, Esophageal Manometry

Venkatasubramanian, Ganapriya, *Arizona State University, Tempe, Arizona*, Functional Electrical Stimulation

Veraart, Claude, *Catholique University of Louvain, Brussels, Belgium*, Visual Prostheses

Verdonck, Pascal, *Ghent University, Belgium*, Hemodynamics

Vermarien, Herman, *Vrije Universiteit Brussel, Brussels, Belgium*, Phonocardiography, Recorders, Graphic

Veves, Aristidis, *Harvard Medical School, Boston, Massachusetts*, Cutaneous Blood Flow, Doppler Measurement of

Vicini, Paolo, *University of Washington, Seattle, Washington*, Pharmacokinetics and Pharmacodynamics

Ville, Jäntti, *Tampere University of Technology, Pori, Finland*, Monitoring in Anesthesia

Vrba, Jini, *VSM MedTech Ltd.*, Biomagnetism

Wagner, Thomas, H., *M. D. Anderson Cancer Center Orlando, Orlando, Florida*, Radiosurgery, Stereotactic

Wahlen, George E., *Veterans Affairs Medical Center and the University of Utah, Salt Lake City, Utah*, Anorectal Manometry

Walker, Glenn M., *North Carolina State University, Raleigh, North Carolina*, Microfluidics

Walterspacher, Dirk, *The Johns Hopkins University, Baltimore, Maryland*, Electroencephalography

Wan, Leo Q., *Liu Ping, Columbia University, New York, New York*, Cartilage and Meniscus, Properties of

Wang, Ge, *University of Iowa, Iowa City, Iowa*, Computed Tomography Simulators

Wang, Haibo, *Louisiana State University Health Center Shreveport, Louisiana*, Monitoring, Umbilical Artery and Vein, Ambulatory Monitoring

Wang, Hong, *Wayne State University, Detroit, Michigan*, Anesthesia, Computers in

Wang, Le Yi, *Wayne State University, Detroit, Michigan*, Anesthesia, Computers in

Wang, Qian, *A & M University Health Science Center, Dallas, Texas*, Strain Gages

Warwick, Warren J., *University of Minnesota Medical School, Minneapolis, Minnesota*, Cystic Fibrosis Sweat Test

Watanabe, Yoichi, *Columbia University Radiation Oncology, New York, New York*, Phantom Materials in Radiology

Waxler, Morris, *Godfrey & Kahn S.C., Madison, Wisconsin*, Codes and Regulations: Medical Devices

Webster, Thomas J., *Purdue University, West Lafayette, Indiana*, Porous Materials for Biological Applications

Wegener, Joachim, *University of Oslo, Oslo, Norway*, Impedance Spectroscopy

Wei, Shyy, *University of Michigan, Ann Arbor, Michigan*, Blood Rheology

Weinmeister, Kent P., *Mayo Clinic Scottsdale*, Spinal Cord Stimulation

Weizer, Alon Z., *Duke University Medical Center, Durham, North Carolina*, Lithotripsy

Weller, Peter, *City University , London, United Kingdom*, Intraaortic Balloon Pump

Wells, Jason, *LSU Medical Centre, Shreveport, Louisiana*, Transcutaneous Electrical Nerve Stimulation (TENS)

Wendelken, Suzanne, *Dartmouth College, Hanover, New Hampshire*, Oxygen Analyzers

Whelan, Harry T., *Medical College of Wisconsin, Milwaukee, Wisconsin*, Hyperbaric Oxygenation

White, Robert, *Memorial Hospital, Regional Newborn Program, South Bend, Indiana*, Incubators, Infant

Williams, Lawrence E., *City of Hope, Duarte, California*, Nuclear Medicine Instrumentation

Wilson, Kerry, *University of Central Florida, Orlando, Florida*, Biosurface Engineering

Winegarden, Neil, *University Health Network Microarray Centre, Toronto, Ontario, Canada*, Microarrays

Wojcikiewicz, Ewa P., *University of Miami Miller School of Medicine, Miami, Florida*, Microscopy, Scanning Force

Wolbarst, Anthony B., *Georgetown Medical School, Washington, DC*, Radiotherapy Treatment Planning, Optimization of

Wolf, Erik, *University of Pittsburgh, Pittsburgh, Pennsylvania*, Mobility Aids

Wood, Andrew, *Swinburne University of Technology, Melbourne, Australia*, Nonionizing Radiation, Biological Effects of

Woodcock, Brian, *University of Michigan, Ann Arbor, Michigan*, Blood, Artificial

Wren, Joakim, *Linköping University, Linköping, Sweden*, Thermocouples

Xiang, Zhou, *Brigham and Women's Hospital, Boston, Massachusetts*, Biocompatibility of Materials

Xuejun, Wen, *Clemson University, Clemson, South Carolina*, Biomaterials, Testing and Structural Properties of

Yan, Zhou, *University of Notre Dame, Notre Dame, Indiana*, Bone Cement, Acrylic

Yannas, Ioannis V., *Massachusetts Institute of Technology*, Skin Tissue Engineering for Regeneration

Yaszemski, Michael J., *Mayo Clinic, College of Medicine, Rochester, Minnesota*, Microscopy, Electron

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
$P\mathrm{o}_2$	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 $T_0 = 273.15\,\mathrm{K}$ 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 SI prefixes, their use should be avoided except for SI unit-multiples for area and volume and nontechnical use of centimeter, as for body and clothing measurement.
10	deka	da[a]	
10^{-1}	deci	d[a]	
10^{-2}	centi	c[a]	
10^{-3}	milli	m	
10^{-6}	micro	μ	
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)	$1.333 \times 10^{2\dagger}$
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)	$0.10^{\dagger}$
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)	$1.0 \times 10^{2\dagger}$
rad	gray (Gy)	$1.0 \times 10^{-2\dagger}$
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)	$1.0^{\dagger}$
stokes (kinematic viscosity)	square meter per second (m^2/s)	$1.0 \times 10^{-4\dagger}$
tex	kilogram per meter (kg/m)	$1.0 \times 10^{-6\dagger}$
ton (long, 2240 pounds)	kilogram (kg)	1.016×10^{3}
ton (metric)	kilogram (kg)	$1.0 \times 10^{3\dagger}$
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)	$0.9144^{\dagger}$

HYDROCEPHALUS, TOOLS FOR DIAGNOSIS AND TREATMENT OF

Sandeep Sood
Anders Eklund
Noam Alperin
University of Illinois at Chicago
Chicago, Illinois

INTRODUCTION

Epidemiology

A congenital form of hydrocephalus occurs in roughly 50 in 100,000 live births (6). Hydrocephalus may also be acquired later in life as a result of a brain tumor, following meningitis, trauma, or intracranial hemorrhage. It has been estimated that prevalence of shunted hydrocephalus is about 40/100,000 population in the United States (7). Untreated hydrocephalus has a poor natural history with a mortality rate of 20–25% and results in severe physical and mental disabilities in survivors (8,9). There has been a significant reduction in mortality and morbidity with use of shunting. However, shunting is associated with a high failure rate; a 40% failure rate occurs within the first year after shunting (10). Advances in the technology have lead to the development of a diverse type of shunt systems to circumvent problems related to long-term shunting, such as obstruction, infection, and overdrainage. Yet, studies done to evaluate these devices have not shown a significant long- or short-term benefit from their use compared with the conventional devices (10). It is estimated that, during the year 2000, the cost associated with shunting exceeded one billion dollars in the United States alone (11). Shunt replacement accounted for 43% of shunt procedures. Endoscopic surgery has provided an alternative strategy in patients with obstructive hydrocephalus. However, limited data in the literature suggest that long-term survival of third ventriculostomy is not significantly superior to that of a shunt (12).

Physiology

The CSF flow in the craniospinal system is influenced by two separate processes: (1) the circulation of the CSF from its formation sites to its absorption sites (i.e., bulk flow) and (2) an oscillatory (back and forth) flow during the cardiac cycle (pulsatile flow). The first process governs the overall volume of CSF and thereby influences intracranial pressure (ICP). The second process, the oscillatory movement of the CSF within the craniospinal compartments, is caused by the pulsatile blood flow entering and leaving the intracranial compartment during the cardiac cycle. These two processes occur over different time scales; circulation and replenishing of CSF occurs over minutes, whereas the time scale of the pulsatile CSF flow is milliseconds.

CSF Circulation. Unlike other organ systems, the brain and the spinal cord are unique in being bathed in a clear fluid called cerebrospinal fluid. The exact role that it plays in maintaining the necessary environment for the functioning of the nervous system is unclear. It has been ascribed a role in providing nutrition, removing excess waste, circulating neurotransmitters, maintaining the necessary electrolyte environment, and acting as a shock absorber against trauma.

The distribution of nutrients, or neurotransmitters, and removal of waste products of metabolism, is an unlikely function of CSF, because these chemicals are present in very low concentrations in the CSF. The main function of CSF is to provide buoyancy to support the brain and act as a cushion against trauma. The normal brain weighs about 1500 g; however, supported by the buoyancy of the CSF, its apparent weight is reduced to about 50 g in the cranium. Support for its role in cushioning the brain and spinal cord against trauma comes from clinical conditions like severe spinal canal stenosis. The CSF cushion around at the site of stenosis is markedly reduced. As a result, spinal cord injury often occurs even with minor trauma as the shock waves are directly transmitted from the bone to the spinal cord.

Cerebrospinal fluid is made through a complex process that occurs in the cells of the choroid plexus, which lines the margin of the four fluid-filled spaces in the brain called the ventricles. First, an ultrafilterate of plasma is formed in the connective tissue surrounding the choroidal capillaries. Next, this is converted into a secretion by carbonic anhydrase enzyme present in the choroids epithelium. The CSF is made at a fairly constant rate of about 10 mL/h. Most of the CSF is made in the choroids plexus of the lateral ventricles. Roughly, 20% of the CSF comes from the ventricular walls. As most CSF is made in the lateral ventricles, it is traditionally believed that the CSF bulk flow occurs from the lateral ventricles to the third ventricle, fourth ventricle, and then through the foramen of Magendie and Lushka into the cerebello-pontine cistern and on to the surface of the brain and spinal cord (Fig. 1). A fifth of the CSF runs down around the spinal cord and then back to the cranial subarachnoid space.

The CSF is absorbed by the cells of the arachnoid granulations (13). These are present in the superior sagittal sinus. The process involves pinocytosis of a small quanta of CSF, on the subarachnoid side of the granulations, and discharge into the blood on the venous side. The process is driven by a pressure difference of at least 5 mm Hg between the subarachnoid CSF and the superior sagittal sinus. A small proportion of CSF is also absorbed along the perivascular spaces and along the nerve sheaths exiting the spinal canal (14).

This traditional view has been recently challenged. Johnston et al. in experimental and cadaveric studies have demonstrated that a large amount of CSF is present around the olfactory nerve and the cribriform plate area

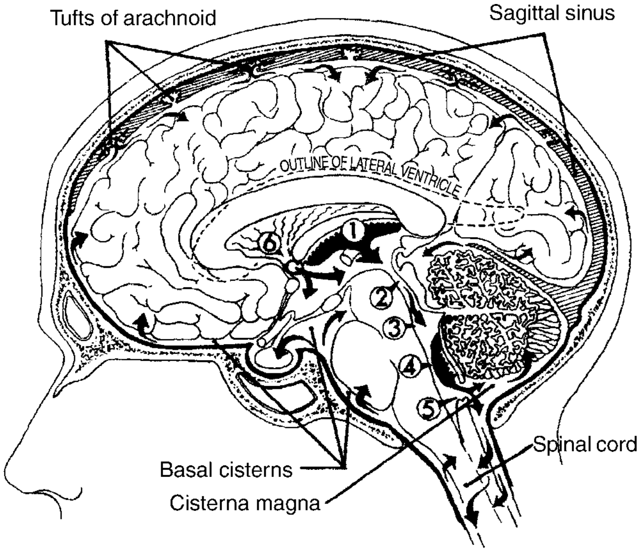


Figure 1. CSF is mainly formed by the choroids plexus in the lateral, third (1), and fourth (4) ventricles. The fluid flows (in the direction of the arrows) from lateral ventricles through the foramen of Monro (6) into the third ventricle. CSF then passes through the aqueduct of Sylvius (2) into the fourth ventricle (3) and exits the fourth ventricle through the foramen of Luschka and Magendie (5) into the cisterna magna and the basal cisterns. The flow is then into the subarachnoid space over the surface of the brain and about the spinal cord. Finally, the fluid is absorbed through the tufts of arachnoid (arachnoid villi) into the sagittal sinus.

and drains into the lymphatic system of the face (15,16). Others believe that CSF may be absorbed directly at the level of the capillaries and perivascular spaces (17).

CSF Pulsations. The pulsatile back and forth movement of CSF between the cranium and the spinal canal with each heartbeat plays a role in modulating the pulsatile cerebral blood flow. Blood flow in the arteries leading blood to the brain is pulsatile, whereas the pulsatility of blood flow in the cerebral veins is considerably attenuated. The displacement of CSF into the spinal canal during the systolic phase helps accommodate the temporary increase in blood volume in the intracranial space, which otherwise has only a limited capacity to accommodate additional volume due to the limited compliance of the intracranial compartment. Reducing the pulsatility of the blood flow through the brain may play a role in diffusion of nutrients to the brain cells from the blood and of waste products from the brain cell to the blood through a less pulsatile flow at the level of the capillaries. As discussed later, MRI measurements of the pulsatile arterial, venous, and CSF flows, to and from the cranium, can now be used to measure intracranial compliance and pressure (ICP), noninvasively. As one of the main roles of shunting is to protect the brain from increased ICP, diagnostic noninvasive measurement of ICP may aid in management decisions in hydrocephalous.

Pathophysiology

Hydrocephalus occurs if there is a mismatch between the CSF production and the absorption. Accumulation of CSF can occur from obstruction to the egress of CSF from the ventricles. This is referred to as obstructive or noncommunicating hydrocephalus. It may also result from impairment of absorption of the CSF at the level of the arachnoid villi or increased resistance to the flow of CSF in the subarachnoid spaces from fibrosis and scarring related to meningitis or previous subarachnoid hemorrhage. This is referred to as communicating hydrocephalus. Irrespective of the cause, the accumulation of CSF has two consequences. It results in an increase in the pressure in the cranium and may cause diltation of the ventricles (ventriculomegaly).

- **Increase in Intracranial Pressure:** Maintaining normal intracranial pressure is important for the functioning of the brain. The pressure in the intracranial cavity increases exponentially with an increase in the total volume of its content (brain tissue, blood, and the CSF) (18). Therefore, increase in intracranial volume, due to uncompensated accumulation of CSF, increases ICP and reduces intracranial compliance. Compliance quantifies the ability of a compartment to accommodate increase in volume for a given increase in pressure and is defined as the ratio of the changes in volume and pressure:

$$\text{Compliance} = \frac{\Delta v}{\Delta p} \tag{1}$$

  where, Δv is change in volume and, Δp is the change in pressure. Intracranial compliance decreases with increased ICP because of the exponential relationship between ICP and intracranial volume (ICV).

Normal ICP is about 1–5 mm Hg in an infant and up to 20 mm Hg in an adult. It is measured by inserting a needle into the spinal canal and recording the pressure using a manometer or by placing a catheter with miniature strain gauge transducer at its distal tip (Codman, Raynham, MA; Camino, Integra LifeSciences, Plainsboro, NJ) directly into the brain parenchyma or the ventricles through a small twist drill hole in the skull. Noninvasive means for measurement of ICP would be important for diagnosis and management of hydrocephalus. Over the last several decades, different approaches have been attempted (19). A method based on measurements of CSF and blood flows to and from the brain by MRI is described in more detail in this article. Increase in ICP can affect the brain in two ways. First, it reduces perfusion of blood into the brain due to the reduced cerebral perfusion pressure (i.e., arterial pressure minus ICP). Depending on the severity and duration, it may result in chronic ischemia causing impairment in higher mental functions, developmental delay in children, or an acute ischemic injury and stroke. Second, rise in pressure in any one of the compartments in the cranium, formed by the tough dural falx in the midline and the tentorium between the cerebral hemispheres superiorly and the cerebellum inferiorly, forces the brain to herniate. This often leads to infarction of the brain stem and death.

- **Symptoms:** Clinically, patients who have elevated ICP generally present with typical symptoms. Headache is the most common. It occurs especially in the early hours of the morning in initial stages. Low respiratory rate during sleep results in buildup of blood CO_2 and vasodiltation. This aggravates the increased ICP in the early stages of the disease. Vomiting is the next common symptom and probably results either from the distortion of the brain stem vomiting center or its ischemia. Vomiting is often associated with retching and rapid respiration that lowers the blood CO_2 level. This in turn leads to vasoconstriction and lowers the ICP and often results in a transient relief in headaches. Diplopia or double vision is also commonly encountered in a setting of increased ICP. It is a result of stretch of the sixth cranial nerve, which controls the abduction of the eyes. Weakness of ocular abduction disturbs the normal axial alignment of the two eyes resulting in defective fusion of the two images by the brain. Blurring of vision and visual loss may occur in patients with long-standing intracranial hypertension. This results from edema of the optic nerve head as the axoplasmic flow in the neurons of the optic nerve is impaired by the high ICP that is transmitted to the nerve through the patent nerve sheath. Hearing deficits related to similar effect on the cochlea are, however, less frequent. Lethargy or sleepiness is frequently observed in patients with high ICP and is probably from a combination of decreased cerebral perfusion and distortion of the brain stem.

- **Ventricular Enlargement:** Depending on the pathophysiology of hydrocephalus, CSF may accumulate only in the ventricles as in obstructive hydrocephalus or in both the ventricles and the subarachnoid space in communicating hydrocephalus. The increased pressure within the ventricle is transmitted to the periventricular region and results, over time, in loss of neurons, increase in periventricular interstitial fluid, and subsequent gliosis with loss of white matter (20).

When onset of hydrocephalus occurs early in infancy, before the skull sutures have closed, the enlarging ventricles are associated with a progressive increase in head circumference and developmental delay. In later childhood and adults, the increasing ventricular size is associated with symptoms of increased ICP. However, ventricular enlargement may also occur with normal mean ICP, in the elderly patients (21). This is referred to as normal pressure hydrocephalus (NPH). The enlarging ventricle stretches the periventricular nerve fibers. The patient presents not with signs of increase in ICP but with progressive gait ataxia, bladder incontinence, and dementia. Similar presentation may also be observed in adolescents with aqueductal stenosis and obstructive hydrocephalus. These patients with compensated long-standing hydrocephalus have been referred to as long-standing hydrocephalus of adults (LOVA) (22).

It is not clear why the ventricles enlarge preferentially, compared with the subarachnoid space, even though the pressure distributes equally in a closed system. It has been argued that, rather than the actual mean ICP, it is the pulse pressure that determines ventricular diltation. Di Rocco et al. (23) have shown that ventricular enlargement could be induced by an intraventricular pulsatile balloon with a high pulse pressure, despite the mean ICP being normal. It may be argued that in a pulsatile system, it is the *root mean square (*RMS) of the pressure, rather than the mean pressure, that is the cause of enlarged ventricles. It has been suggested that, in communicating hydrocephalus, decrease in compliance may be responsible for preferential transmission of the pulsations to the ventricles (24,25). However, others have shown that in acute or chronic communicating hydrocephalus, the pulse pressure and the pressure waveforms in the SAS and the ventricles are similar (26,27). Alternative explanations offered are that the pia over the cortical surface is more resilient than ependyma that lines the ventricular wall; the venous pressure in the periventricular region is lower, making it more deformable than the subcortical area (28).

DIAGNOSTIC METHODS

Measurement of Resistance to CSF Reabsorption

The hydrodynamics of the craniospinal system is governed by patient-specific properties like CSF formation rate, CSF reabsorption resistance (historically termed as outflow resistance), venous pressure in the sinus, and craniospinal compliance. Together with the periodic variations in ICP, due to blood volume variation from the heartbeat and vasomotion, these properties describe the CSF dynamics, which provide the working environment of the brain. When this environment is disturbed, it affects the function of the brain resulting in the clinical symptoms of hydrocephalus. After shunting, symptoms are often eliminated or reduced. It shows that a clinical improvement can be accomplished by actively changing the brain's working environment. This link among CSF dynamics, brain function, symptoms, and shunting has made researchers look for CSF dynamical means to identify patients that would benefit from a shunt surgery. Outflow resistance has been suggested as a strong predictive parameter in communicating hydrocephalus. Invasive infusion tests in conjunction with a mathematical model of the craniospinal system can be used to estimate CSF absorption rate. The most accepted model for the system hydrodynamics has been proposed by Marmarou (29).

The basic assumptions for the model are as follows:

- CSF reabsorption rate is linearly dependent on the difference between the intracranial and venous pressures (the outflow resistance describes this linear relationship)
- A pressure-dependent compliance
- A constant formation rate of CSF, independent of ICP

The model can be displayed as an electrical analogy (Fig. 2). The model is described mathematically as a differential equation of the time-dependent ICP as a function of external infusion and the governing physical parameters:

$$\frac{dP_{\mathrm{IC}}(t)}{dt} + \frac{K}{R_{\mathrm{out}}}[P_{\mathrm{IC}}(t)]^2 - \left(K \cdot I_{\mathrm{infusion}}(t) + \frac{K \cdot P_{\mathrm{r}}}{R_{\mathrm{out}}}\right) P_{\mathrm{IC}}(t) = 0 \qquad (2)$$

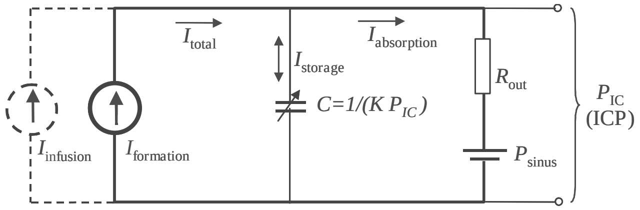


Figure 2. Model of the dynamics of the CSF system. $I_{\text{formation}}$ is the CSF formation rate, C is the pressure-dependent compliance described by the elastance parameter K, R_{out} is outflow resistance, P_{sinus} is the venous pressure in the sinus, and P_{IC} is the intracranial pressure. I_{infusion} is the option of external infusion of artificial CSF.

where R_{out} is outflow resistance, P_{IC} is the intracranial pressure, P_{r} is the ICP at rest, and K is the elastance.

Estimation of CSF outflow resistance and the other system parameters requires perturbation of the system steady state by infusion of fluid into the craniospinal system, either through a lumbar or a ventricular route. Typically, one or two needles are placed in the lumbar canal. When two needles are used, one is connected to a pressure transducer for continuous recording of the dynamic changes in ICP following the infusion, and the other one for infusion and/or withdrawal of the fluid. Different protocols of infusion will lead to unique mathematical solutions. In addition to the resting pressure (also refereed to as opening pressure), which always is determined during these investigations, it is generally believed that the outflow resistance is the clinically most important parameter, but compliance has also been proposed as a predictor for outcome after shunting.

Bolus Infusion. An example of ICP recoding during the bolus infusion test is shown in Fig. 3. The approach is to first determine the compliance from the ratio of the injected volume and the magnitude of pressure increase (30). A pressure volume index (PVI), which describes compliance, is calculated through the expression:

$$\text{PVI} = \frac{\Delta V}{\log(P_{\text{p}}/P_0)} \tag{3}$$

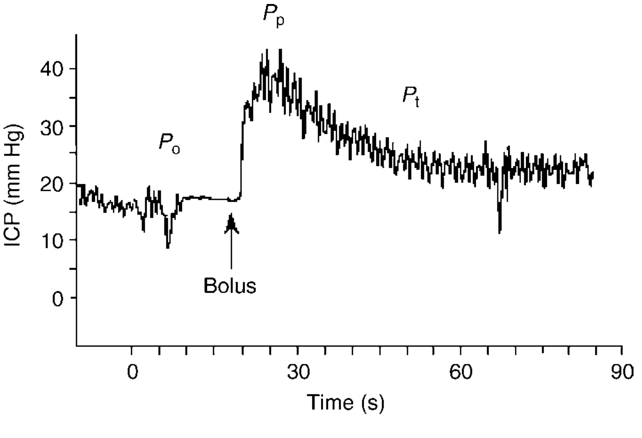


Figure 3. ICP curve from a bolus injection of 4 mL. Figure from Marmarou et al. (30).

where ΔV is the infused volume, P_{p} is the peak pressure and P_0 is the initial pressure just before the infusion. The next step is to determine R_{out} from the spontaneous relaxation curve when the ICP returns toward the resting pressure (Fig. 3). Solving the differential equation for the relaxation phase after the bolus infusion gives the following expression for R_{out} as a function of time (31):

$$R_{\text{out}} = \frac{tP_0}{\text{PVI}\ \log\left[\dfrac{(P_t/P_p)(P_p - P_0)}{P_t - P_0}\right]} \tag{4}$$

where t is the time in seconds after the bolus and P_t is the measured pressure at time t on the relaxation curve. From each bolus, a number of values of R_{out} are calculated and averaged, for example at $t = 1$ min, 1.5 min, and 2 min. The bolus procedure is usually repeated a couple of times for increased measurement reliability.

Constant Pressure Infusion. In this infusion protocol, several constant ICP levels are created. This is done by using a measurement system that continuously records the ICP and regulates it by controlling the pump speed of an infusion pump (Fig. 4) (32). The net infusion rate needed to sustain ICP at each pressure level is determined, and a flow versus pressure curve is generated (Fig. 4). Using linear regression, the outflow resistance is then determined from the slope of that curve (33), because at steady state, the differential equation reduces to

$$I_{\text{inf}} = \frac{1}{R_{\text{out}}} P_{\text{IC}} - \frac{P_{\text{r}}}{R_{\text{out}}} \tag{5}$$

where P_{IC} is the mean intracranial pressure on each level, P_{r} is the resting pressure, and I_{inf} is the net infusion flow at each level.

The constant pressure method can also be used to estimate the CSF formation rate. This is done by lowering the ICP beneath the venous pressure, i.e., below 5 mm Hg. At that ICP, no CSF reabsorption should take place. Therefore, the net withdrawal of CSF needed to sustain that constant pressure level should equal the formation rate.

Constant Flow Infusion. In this method, both the static and the dynamic behavior of the CSF system can be used to estimate outflow resistance (34). In a steady-state analysis R_{out} can be calculated from the stable ICP value associated with a certain constant infusion rate. R_{out} is then estimated by the following expression:

$$R_{\text{out,stat}} = \frac{P_{\text{level}} - P_{\text{r}}}{I_{\text{inf}}} \tag{6}$$

where $R_{\text{out,stat}}$ is a static estimation of R_{out}, P_{level} is the new equilibrium pressure obtained at the constant infusion rate, P_{r} is the resting pressure, and I_{inf} is the infusion rate (Fig. 5).

R_{out} can also be estimated from the dynamic phase during the pressure increases toward the new equilibrium (Fig. 5). This procedure will also give an estimate of the craniospinal compliance (elastance). The differential equation is now solved for a condition of a constant infusion rate, and the solution is fitted against the recorded pressure

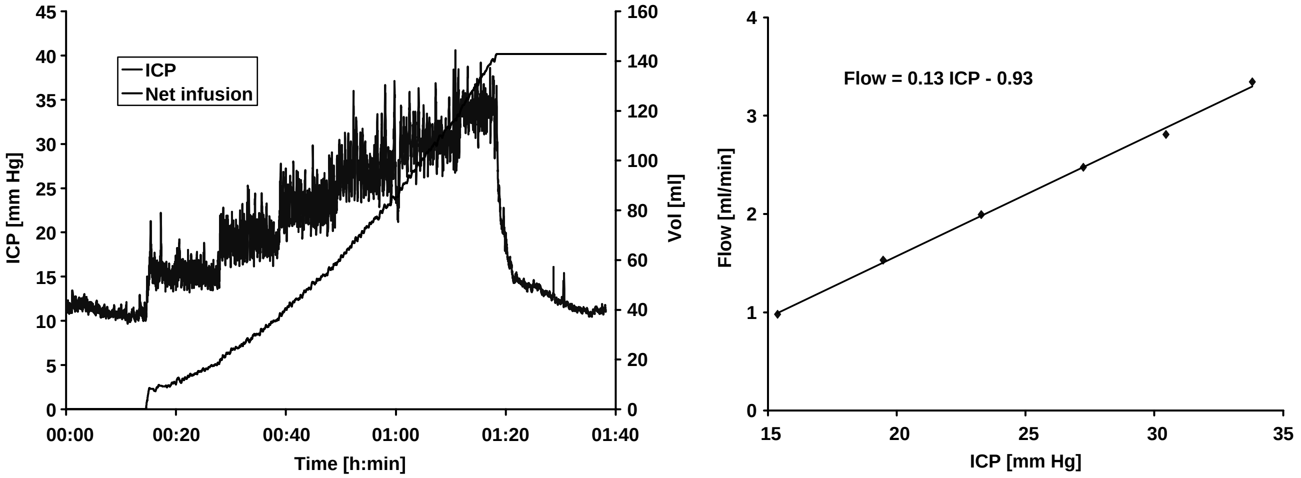


Figure 4. Pressure and flow curves for the constant pressure method. Left graph shows typical data for ICP and infused CSF volume versus time. Right graph shows the mean pressure and flow points determined from each steady-state level. Outflow resistance corresponds to the inverse of the slope.

curve. The time-dependent pressure increase is described by the following expression:

$$P(t) = \frac{\left(I_{\text{inf}} + \frac{P_{\text{r}} - P_0}{R_{\text{out,dyn}}}\right) \cdot (P_{\text{r}} - P_0)}{\frac{P_{\text{r}} - P_0}{R_{\text{out,dyn}}} + I_{\text{inf}} \left[e^{-K\left(\frac{P_{\text{r}} - P_0}{R_{\text{out,dyn}}} + I_{\text{inf}}\right)t} \right]} + P_0 \qquad (7)$$

where K is the elastance and P_0 is a reference pressure that is suggested to be equal to venous sinus pressure. Fitting against data will result in estimations of the unknown parameters $R_{\text{out,dyn,}}$ K, and P_0.

In summary, CSF infusion tests are conducted to reveal parameters describing the hydrodynamics of the craniospinal system. An active infusion of artificial CSF is performed, and the resulting ICP response is recorded, and parameters such as outflow resistance, compliance, formation rate, and the venous sinus pressure are then estimated based on a proposed mathematical model. Outflow resistance values determined with the bolus method are usually lower than the values determined with the constant infusion and constant pressure methods. The reason for this difference is not well understood at this time. Determination of R_{out} is often used as a predictive test in hydrocephalus, and it has been stated that if the outflow resistance exceeded a certain threshold, it is an excellent predictor of clinical improvement after shunting (35). In a recent guideline for idiopathic normal pressure hydrocephalus, measurement of R_{out} is included as a supplementary test for selecting patients suitable for shunt surgery (36).

DIAGNOSIS WITH IMAGING

Cross-sectional imaging is routinely used in the diagnosis of hydrocephalous. CSF spaces are well visualized with CT and MRI. In CT images, CSF spaces appear darker due to the lower atomic density of the CSF compared with that of brain tissue. MRI provides an excellent soft-tissue contrast resolution and is considered the primary imaging modality for brain imaging. With MRI, CSF spaces can appear either darker or brighter compared with its surrounding tissues depending on the imaging technique. An example of a CT image and MRI images demonstrating abnormally large CSF spaces is shown in Fig. 6. Cross-sectional imaging enables quantitative assessment of the CSF spaces as well

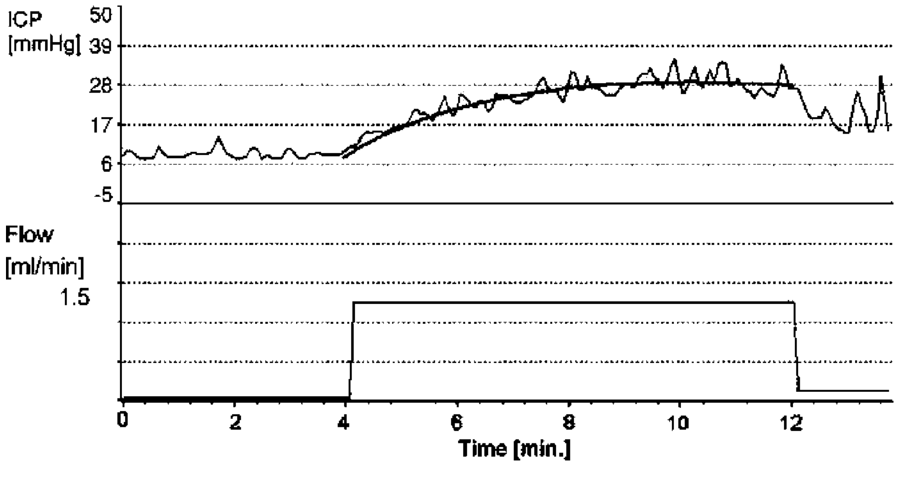


Figure 5. ICP data from a constant infusion investigation. Figure modified from Czosnyka et al. (34).

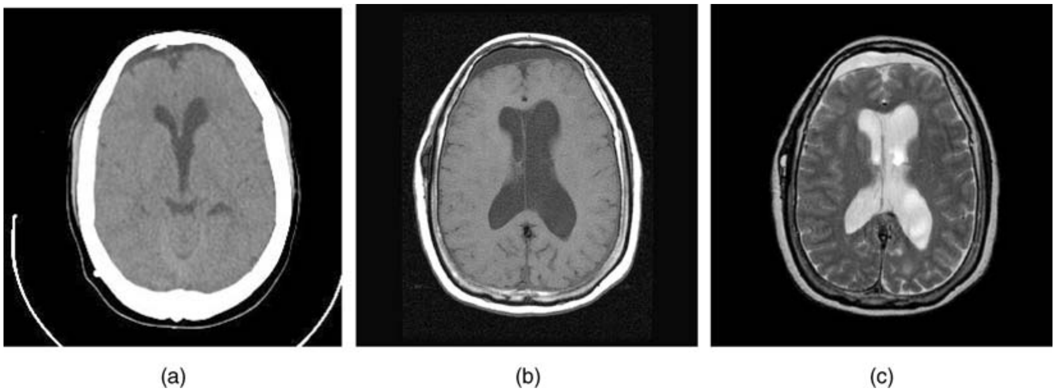


Figure 6. An example of a (a) CT image and (b and c) MRI images demonstrating abnormally large CSF spaces. The appearance of CSF in MRI depends on the technique used to acquire the images; its (b) dark with a T1 technique and (c) bright with a T2 technique.

as 3D reconstruction of the geometry of the ventricular system. The 3D model is obtained by segmentation of the CSF spaces in each of the 2D slices. An example of 3D models of the ventricular system from MRI data demonstrating normal size ventricles and enlarged ventricles are shown in Fig. 7a and b, respectively.

MRI-based motion-sensitive techniques capable of imaging flow are gaining an important role in the diagnosis of hydrocephalus. In particular, dynamic phase-contrast techniques provide images of velocities (velocity-encoded images). The degree of brightness in these images is proportional to the direction and the speed of the moving fluid or tissue. Dynamic (cine) phase contrast images are used to visualize the back and forth flow through the different CSF pathways. The cine phase contrast MRI (PCMRI) technique is also used to derive quantitative parameters such as CSF volumetric flow rate through the aqueduct of Sylvius, from which the CSF production rate in the lateral ventricles can be estimated (37), and intracranial compliance and pressure (19,30).

MRI-Based Measurement of Intracranial Compliance and Pressure

The noninvasive measurement of compliance and pressure uses the cardiac pulsations of the intracranial volume and pressure (30,38). This method is the noninvasive analogs to the measurement of intracranial compliance with the previously described bolus infusion method where the volume and pressure changes are calculated from the MRI measurements of CSF and blood flows to and from the brain. Intracranial elastance, i.e., a change in pressure due to a small change in volume, or the inverse of compliance, is derived from the ratio of the magnitudes of the changes in volume and pressure, and the pressure is then derived through the linear relationship between elastance and pressure. The MRI method measures the arterial, venous, and CSF flows into and out of the cranial vault. A small-volume change, on the order of 1 mL, is calculated from the momentary differences between inflow and outflow at each time points in the cardiac cycle. The pressure change is proportional to the pressure gradient change, which is calculated from time and spatial derivatives of the CSF velocities using fluid dynamics principles.

A motion-sensitive MRI technique, cine phase contrast, provides a series of images where the value at each picture element is proportional to the velocity at that location. The phase contrast MRI technique is based on the principle that the precession frequency of the protons is proportional to the magnetic field strength. Therefore, velocity can be phased-encoded by varying the magnetic field in space and time, i.e., generating magnetic field

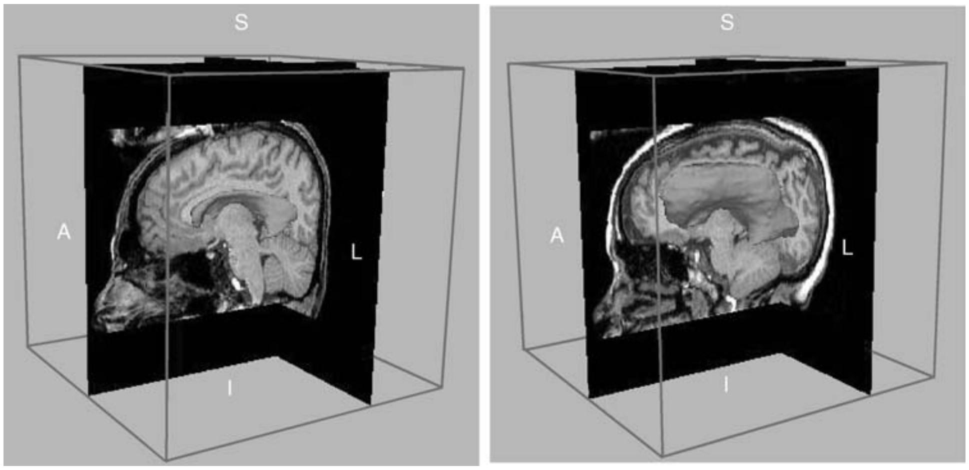


Figure 7. Volume rendering of the CSF spaces inside the brain (i.e., ventricles) generated using segmented MRI data from a (left) healthy volunteer and from a (right) hydrocephalic patient.

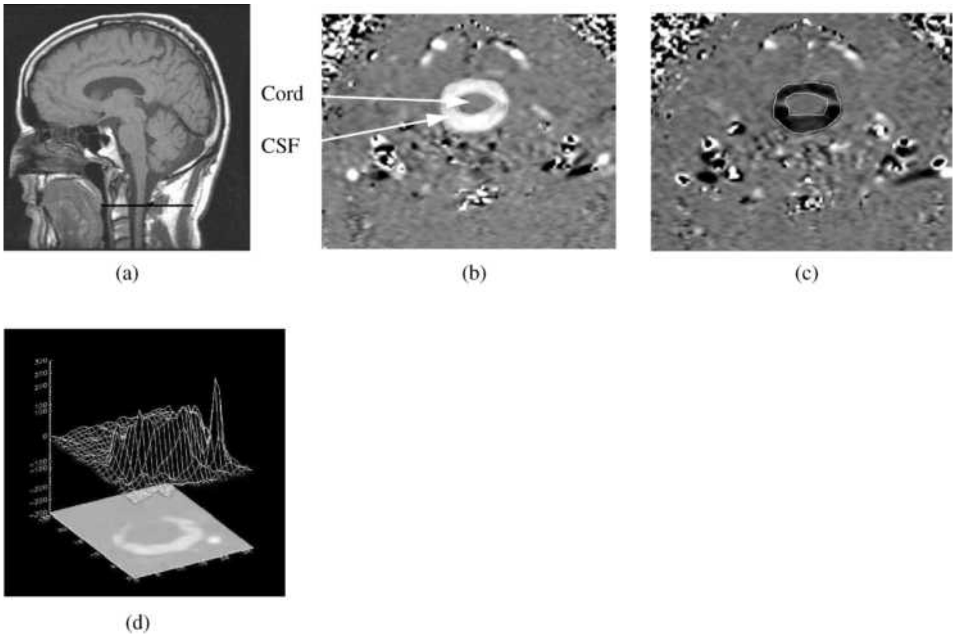


Figure 8. (a) Anatomical mid-sagittal T1-weighted MR image showing the location of the axial plane used for CSF flow measure- ment (dark line). (b and c) Phase-contrast MRI images of CSF flow in the spinal canal. (b) CSF flow during systole. (c) CSF flow during diastole. The pixel values in these images are proportional to velocities in a direction perpendicular to the image plane. Gray-static tissue, white-outward flow (caudal direction), and black-inward flow (cranial direction). (d) A 3D plot of the CSF velocities during systole.

gradients. When a gradient field is applied along an axis for a short time, the proton's phase will change based on its location along that axis. When a bipolar (positive and then negative) gradient field is applied, the phase of the stationary protons will increase during the positive portion (lobe) of the bipolar gradient and then will decrease during the negative lobe. If the lobes were of equal area, no net phase change would occur. However, moving protons, such as those in the blood or CSF, will experience different field strength during each lobe due to their change in position; this will result in a net phase change proportional to the proton velocity.

Examples of MRI phase contrast images of CSF and blood flow are shown in Figs. 8 and 9, respectively. The oscillatory CSF flow between the cranial and the spinal compartments is visualized in images taken in a transverse anatomical orientation through the upper cervical spinal canal. The location of this plane is indicated on a mid-sagittal scout MR image shown in Fig. 8a. Fig. 8b depicts outflow (white pixels) during systole, and Fig. 8c depicts

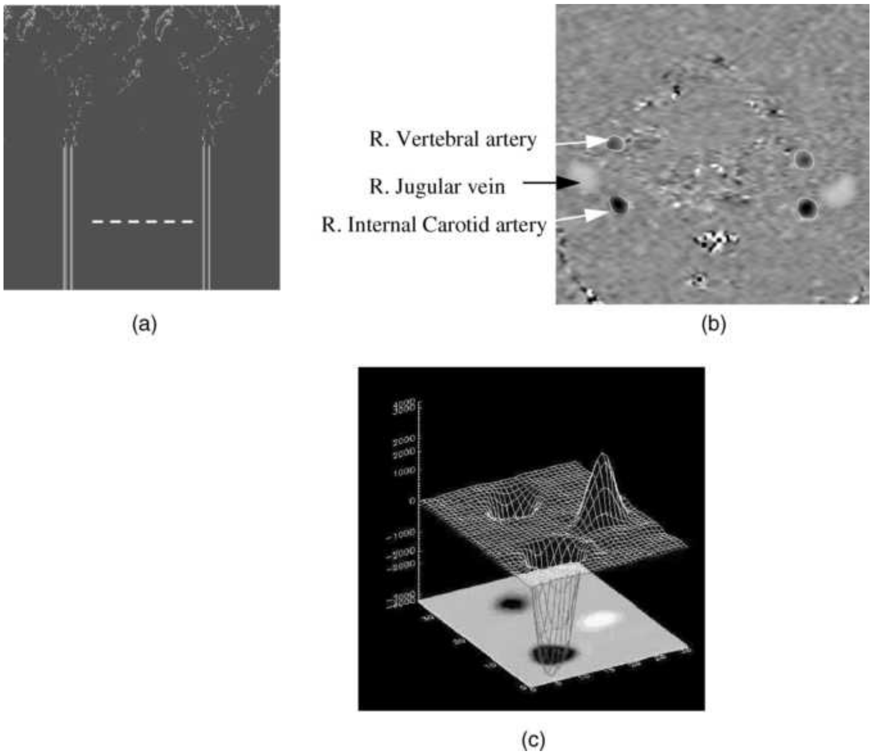


Figure 9. (a) A blood vessel MRI scout image showing the location of the axial plane for blood flow measurement (dash line). (b) A phase contrast MRI image of blood flow through that location. Black pixels indicate arterial inflow, and white are venous outflow. (c) A 3D plot of the blood flow velocities.

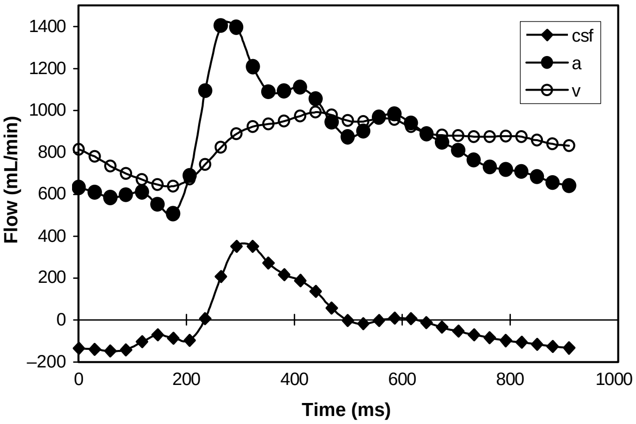


Figure 10. The volumetric flows into and out of the cranial vault during the cardiac cycle derived from the MRI phase contrast scans. Total arterial inflow (filled circles), venous outflow (open), and the cranial-to-spinal CSF volumetric flow rate (diamonds) during the cardiac cycle. Note that arterial inflow is greater than venous outflow during systole.

inflow (black pixels) during diastole. Fig 9d depicts a 3D plot of the velocities in a region of interest containing the CSF space and an epidural vein. The CSF flow is imaged with a low-velocity encoding, and the faster blood flow through the neck arteries and veins is imaged using high-velocity encoding. The location of the imaging plane used for blood flow measurement is shown in Fig. 9a, and a velocity encoded image of blood flow is shown in Fig. 9b. Fig 9c depicts a 3D plot of the velocities in a region of interest containing the internal carotid and vertebral arteries and the jugular vein.

Volumetric flow rates are obtained by integration of the velocities throughout a lumen cross-sectional area. The total volumetric arterial flow rate—that is, total cerebral blood flow—is calculated directly from the sum of the volumetric flow through the four vessels carrying blood to the brain (internal carotid and vertebral arteries). The venous blood outflow is obtained by summation of the flow through the jugular veins, and through secondary venous outflow channels such as the epidural, vertebral, and deep cervical veins when venous drainage occurs through these veins. An example of the volumetric flow waveforms for CSF, arterial inflow, and venous outflow measured in a healthy volunteer is shown in Fig. 10.

The rate of the time-varying intracranial volume change (net transcranial volumetric flow rate) is obtained by subtracting outflow rates from inflow rates at each time point. The intracranial volume change (delta of volume from a given reference point) is obtained by integrating that waveform with respect to time. Waveforms of the net transcranial volumetric flow rate and the change in the intracranial volume are shown in Fig. 11.

The magnitude of the change in intracranial pressure during the cardiac cycle (pulse pressure) is proportional to that of the CSF pressure gradient waveform. A method to measure pressure gradient of pulsatile flow in tubes with MRI was reported by Urchuk and Plewes (39). Pulsatile pressure gradients are derived from the MRI velocity-encoded phase contrast images using the Navier–Stokes relationship between pressure gradient and temporal and spatial derivatives of the fluid velocity for incompressible fluid in a rigid tube (40). Pressure traces from invasive recordings obtained invasively in patients with low and elevated ICP with an intraventricular pressure transducer and the corresponding CSF pressure gradient waveforms derived from the MRI measurements of the CSF velocities at low- and high-pressure states are shown in Fig 12. The ratio of the magnitude of the pressure and volume changes, i.e., intracranial elastance, is then expressed in terms of MR-ICP based on the linear relationship between elastance and ICP.

DEVICES FOR TREATMENT

Despite significant advances in understanding of the pathophysiology of hydrocephalus, the gold standard for the treatment of hydrocephalus still continues to be CSF diversion through a tube shunt to another body cavity. Unfortunately, treatment with CSF shunts is associated with multiple complications and morbidity. The rate of shunt malfunction in the first year of shunt placement is 40%, and, thereafter, about 10% per year. The cumulative risk of infection approaches 20% per person although the risk of infection per procedure is only 5–8% (41). The technological advances in shunt valve designs and materials have had only a marginal impact on the rate of complications. Third ventriculostomy has become popular in recent years for management of obstructive hydrocephalus, but many questions about its long-term permanence remain controversial. Choroid plexectomy (42,43) aimed at arresting hydrocephalus by reducing CSF production or pharmacotherapy with similar intentions have had very limited success in selected patients.

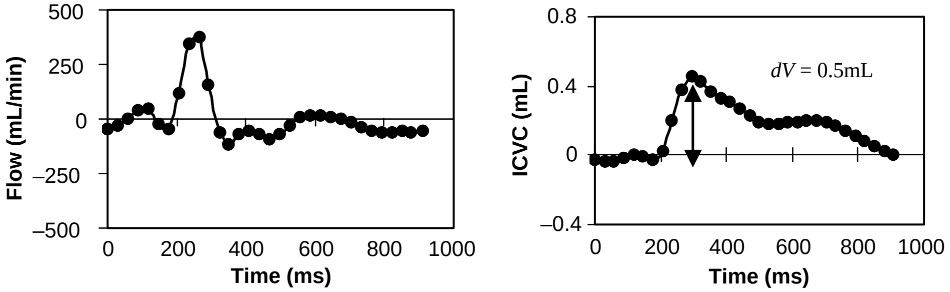


Figure 11. (Left) The MRI-derived net transcranial volumetric flow rate waveform. (Right) The intra cranial volume change during the cardiac cycle derived by integrating the net transcranial volumetric flow waveform on the left. Note that the maximal volume change in this subject is 0.5 mL.

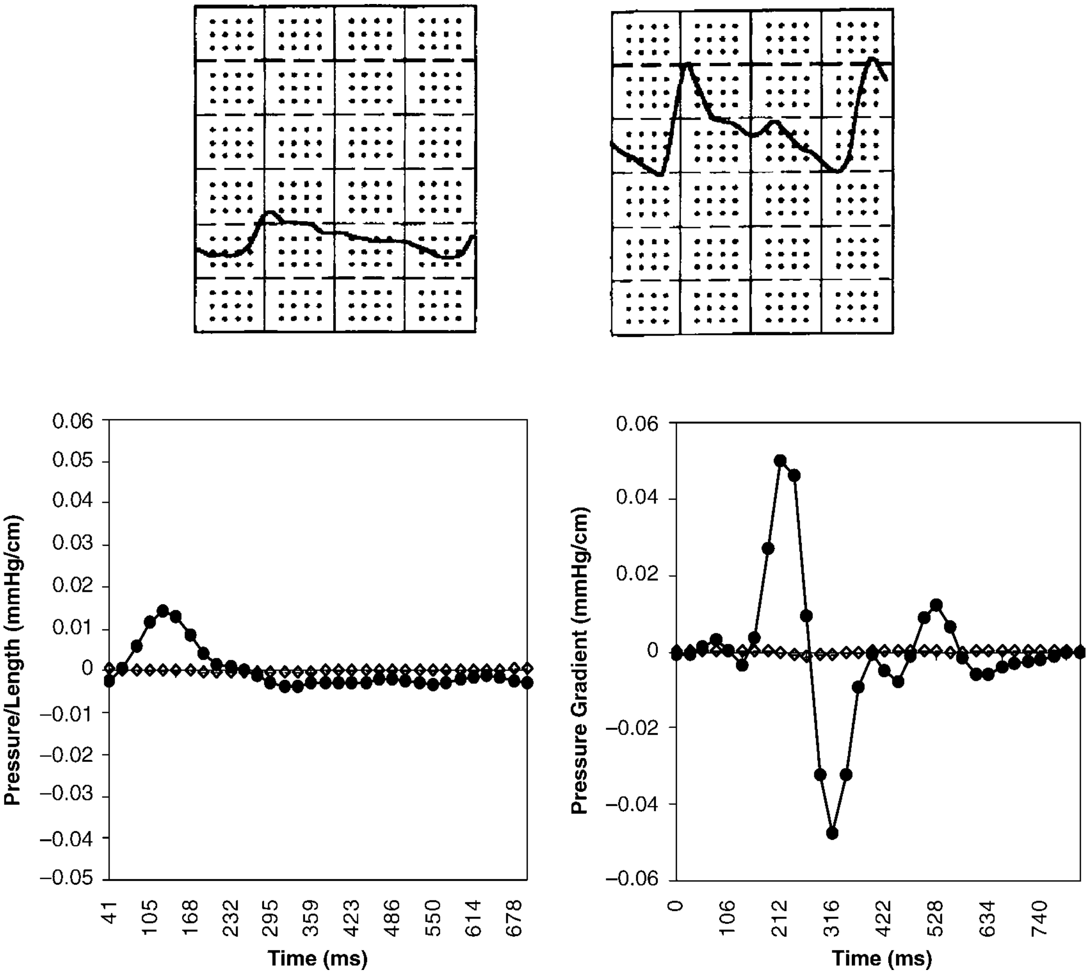


Figure 12. Invasive pressure traces (top) obtained with an intra-ventricular catheter from two patients with (left) low and (right) elevated ICP. The corresponding MRI-derived CSF pressure gradients are shown at the bottom.

Nonobstructive Hydrocephalus

No treatment other than CSF diversion has been effective in management of this form of hydrocephalus. The CSF may be diverted from the ventricles through a catheter that runs in the subcutaneous tissue into the abdominal cavity where it is absorbed by the peritoneum (*ventriculo- peritoeal shunt*) (Fig. 13). It may also be diverted from the spinal subarachnoid space by a lumbar shunt that diverts it to the peritoneum (*Lumbar-peritoneal shunt*). Lumbar CSF diversion avoids the potential risk of brain injury by the ventricular catheter. Lumbar shunts have a lower risk of obstruction and infection (44) but are more prone to malfunction from mechanical failures (45), and, the development of hind brain herniation, over a period of time, has been well documented (46,47). Evaluation of a lumbar shunt for function is more cumbersome than that of a ventricular shunt. The lumbar shunt is usable in patients with communicating hydrocephalus, small ventricles, and patients who have had multiple ventricular shunt malfunctions.

In patients who cannot absorb CSF from the peritoneum due to scarring from previous operations or infections, the CSF may be diverted to the venous system through a catheter placed at the junction of superior vena cava and the right atrium (*ventriculo / lumbar-atrial shunt*).

A typical shunt system consists of three parts (Fig. 14). First, the proximal catheter, i.e, the catheter, is inserted into the ventricle or the lumbar subarachnoid space. Second, the valve controls the amount of CSF that flows through the shunt system, and third, the distal catheter drains the CSF from the valve to the peritoneum or the atrium.

Proximal Catheter

Three basic types of proximal catheter designs are available: simple with multiple perforations (Codman, Raynham, MA; PS Medical, Goleta, CA), simple Flanged (Heyer-Schulte), Integra, Plainsboro, NJ; Anti-Blok (Phoenix Vygon Neuro, Valley Forge, PA) with receded perforations. The last two have been designed to minimize the growth of choroid plexus into the perforations and causing obstruction. There is no controlled study to suggest that these two designs are in any way superior to simple perforations. The flanged catheters can get stuck, as choroid plexus grows around it, making removal of an obstructed catheter difficult (48).

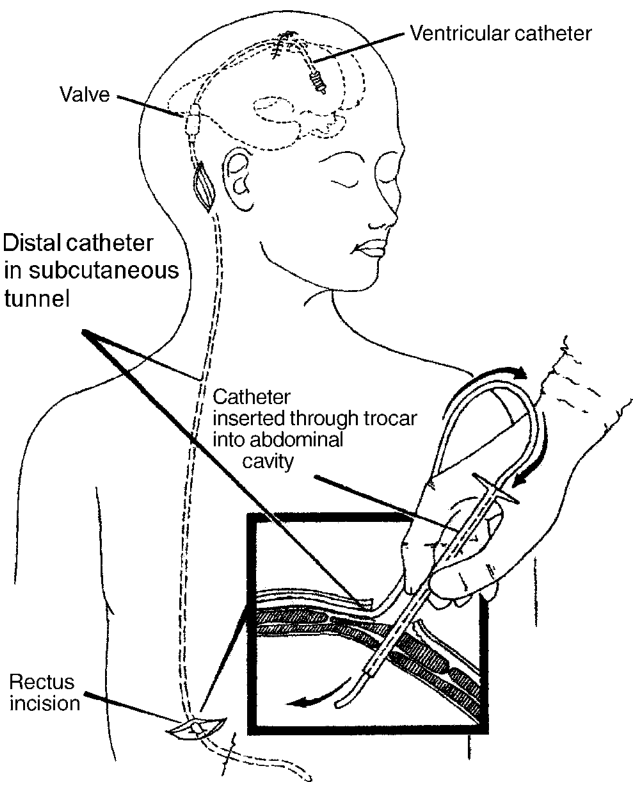

Ventricular catheter
Valve
Distal catheter in subcutaneous tunnel
Catheter inserted through trocar into abdominal cavity
Rectus incision

Figure 13. The shunt consists of three parts. The ventricular catheter enters the skull through a burr hole in the skull and passes through the brain into the lateral ventricle. It is connected to the valve that is placed in the subcutaneous tissue of the scalp. The valve in turn is connected to the distal catheter that runs in the subcutaneous tissue to enter the peritoneal cavity of the abdomen as shown in the inset (*ventriculo-peritoneal shunt*) or into the jugular vein and through it to the superior vena cava (*ventriculo-atrial shunt*).

Placement of the proximal catheter has generated considerable controversy in the literature (48–51). More recently, endoscopic placement of the proximal catheter into the frontal horn, away from the choroid plexus, has been advocated to minimize proximal malfunction (3,52,53). Again no controlled study has been done to confirm whether placement of the proximal catheter into frontal or occipital horn is superior to placement in the body of the lateral ventricle. Often catheters that are grossly malpositioned may continue to work, whereas those that are well positioned may fail. The choice of the site, frontal or parietal, may be made on the basis of the above although some studies have suggested a higher incidence of seizure with catheters placed via a frontal burr-hole (49). A study to evaluate use of endoscope to place the shunt catheter in the frontal horn failed to show any benefit (54). This suggests that no matter where the catheter is placed, the flow of CSF toward the catheter causes the choroids plexus to creep toward the catheter, ultimately causing ingrowth and obstruction of the catheter (55).

To remove an obstructed catheter, intraluminal coagulation of the choroid plexus is done using a stylet and low-voltage diathermy, at the time of shunt revision (56–58). Massive intraventricular hemorrhage may occur if the choroid plexus is torn while forcefully removing the catheter. Delayed subarachnoid hemorrhage from rupture of pseudoaneurysm resulting from diathermy of a catheter close to anterior cerebral artery has been reported (59). At times, if the ventricular catheter is severely stuck, it is advisable to leave it in position but occlude it by a ligature and clip. This may become necessary as sometimes an occluded catheter may become unstuck over time and begin to partially function, resulting in formation of subgaleal CSF collection. Replacing a new catheter into the ventricle in patients with small or collapsed ventricles can be sometimes challenging. In most instances, after removal of the old catheter, the new catheter can be gently passed into the ventricle through the same tract. Frameless stereotaxis (StealthStation, Medtronics, Goleta, PA) is now available and may offer an alternative to cumbersome and time-consuming frame-based stereotactic catheter placement (52).

Valve

The valve regulates the amount of CSF that is drained. The aim is to maintain normal ICP. The simplest valves are differential pressure valves. The CSF drainage in these valves is based on the pressure difference between the proximal and the distal ends. Three major configurations are available (Fig. 14): diaphragm, slit valve, and ball-spring mechanism in different pressure ranges (low, medium, and high). Recently, valves in which the pressure setting can be changed with a magnetic wand have become available. These programmable valves allow pressure changes over different pressure ranges based on the manufacturer. The pressure setting on the valve can be ascertained by X ray of the head in the Medos valve (Codman, Raynham, MA) or using a magnetic wand in the Strata valve (Medtronics, Goleta, CA). To prevent inadvertent changes in the valve setting by stray magnetic fields, the Polaris valve (Sophysa, Costa Mesa, CA) has an ingenious locking mechanism that allows changes only if the magnetic field has a certain configuration.

Slit valves tend to be the most inaccurate in their performance followed by ball and spring valves. The diaphragm valves proved to be most stable in long-term tests. Most valves, like the slit valves, ball–spring, and diaphragm valves, offer a lower resistance (<2.5 mm Hg/mL/min) than the normal physiological CSF outflow of 6–10 mm Hg/mL/min. The standard distal tubing of 110 cm increases the overall resistance to 50–80% of the physiological value (60).

Standard differential pressure valves are available in different pressure ranges. It is unclear whether it makes a difference in an ambulatory patient to use a low-, medium-, or high-pressure valve because in the upright position irrespective of the rating the hydrostatic column converts all differential pressure valves into "negative" pressure valves (61). The overdrainage results in persistent headaches from low ICP, ventricular collapse, and increased risk of shunt obstruction. Long-term changes in cerebrovenous physiology cause acute and severe increase in ICP without enlargement of ventricles at the time of shunt malfunction (62). To circumvent the overdrainage in the

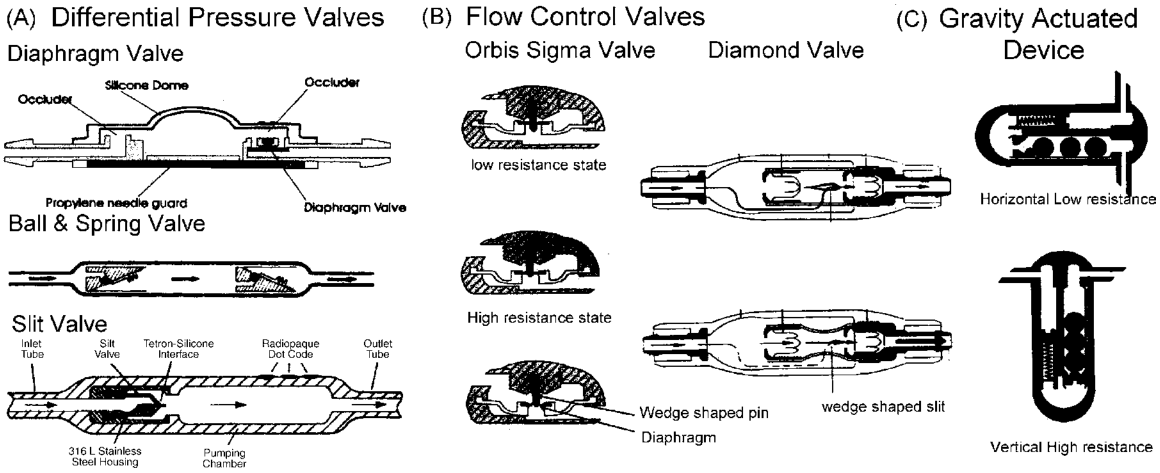


Figure 14. Three major types of valve designs are available. (a) Differential pressure valves allow flow in proportion to the pressure difference between the proximal catheter and the distal catheter. Configurations are a simple diaphragm, ball and spring, or slit valve. Programmable differential pressure valves can be programmed to a pressure setting using a magnetic wand. In the upright position, due to a negative pressure from the hydrostatic column of fluid in the distal catheter, these valves tend to overdrain causing negative pressure symptoms. (b) Flow control valves have the ability to limit overdrainage from a negative hydrostatic pressure gradient. The Orbis-Sigma Valve has a wedge-shaped pin over which the orifice of the diaphragm (arrows) rests. When the distal pressure becomes negative in the upright position, the diaphragm slides downward on the pin narrowing the drainage channel and hence reducing the flow rate. The Diamond valve has a wedge-shaped slit in the construct (arrows) that narrows as the distal pressure becomes increasingly negative again reducing the flow rate. (c) Gravity actuated devices reduce drainage in the upright position by increase in resistance to flow of CSF from the weight of metal balls in the drainage channel.

upright position, ingenious devices, also referred to as devices for reducing siphoning (DRS), have been developed (Fig. 14). The Anti-Siphon device (Integra LifeSciences) has a flexible diaphragm that mechanically senses atmospheric pressure and shuts off the drainage channel if the hydrostatic pressure in the fluid column becomes negative. Flow control valves (Orbis Sigma Valve, Integra LifeSciences and Diamond Valve, Vygon Neuro, Valley Forge, PA) have a drainage channel that narrows as the differential pressure increases in the upright position to reduce the flow. Gravity actuated devices (Gravity Compensating Accessory, Integra LifeSciences, CA, Chabbra Shunt) have metal balls that fall over one another in the upright position to increase resistance to flow. Double channel devices (Dual Switch Valve, Christoph Miethke GmBH & Co KG; SiphonGuard, Codman) have two channels; the low-resistance channel is shut off by a gravity actuated ball in the upright position.

There is no evidence to suggest that use of one type of valve is superior to the other, and several valve designs are available in the market today. A recent multicentric study, evaluating three basis types of valves, failed to confirm the utility of flow control or anti-siphon valves in children and infants over the differential pressure valves (10). Similarly, studies have failed to show that programmable devices are superior to fixed pressure valves (63). Over a period of time, the ventricle tended to become small irrespective of the type of valve used. The rate of proximal malfunction in a patient with flow control valves was 6.5% compared with 42–46% for the other two valves, although the overall rate of malfunction and shunt survival was not statistically different. The design of the flow control valves with a narrow orifice makes it sensitive to malfunction (64). Certainly, revising a valve has less morbidity and risk of neurological injury than revising the proximal catheter, especially in patients with slit ventricles. There is evidence that a significant number of patients do not tolerate flow control valves and, despite a radiologically functioning shunt, have high intracranial pressure from underdrainage through the valve. In patients with limited pressure–volume compensatory reserve, there can be an excessive increase in intracranial pressure during cardiovascular fluctuations, especially at night and be responsible for nighttime or early morning headaches, in patients with flow control devices (60). Self-adjusting diaphragm valves like the Orbis-Sigma (Integra LifeSciences), on bench test, have proved to be inaccurate and unstable at perfusion rates of 20–30 mL/h, which is the most important physiological range, leading to pre-valve pressures rapidly changing between 4 and 28 mm Hg. During long-term perfusion, these may resemble ICP pressure waves (60).

Diaphragm-based anti-siphon devices are prone to obstruction from encapsulation as has been shown in experimental animals and is often encountered in patients who have had recurrent malfunctions (65). Some patients are more prone to develop heavy scarring around the shunt system. Again, there is no evidence that using an open (ASD, Anti Siphon device, Integra LifeSciences) has any advantage over using a closed system that opens when the pressure exceeds the negative hydrostatic pressure (SCD, Siphon Control Device, Medtronics), although theoretically malfunctions in an open system would only result in loss of anti-siphon function without obstruction to the flow of CSF. In the open system (ASD), the flow through the valve stops only after the intracranial pressure has become negative in the upright position, which is more physiological, than with SCD, in which the flow stops once the pressure reaches zero. In the multicentric shunt study, the incidence of overdrainage was 7.8% in the SCD group and 2.6% in the Standard valve group. The study suggests that diaphragm-based anti-siphon devices may not be any superior to differential pressure valves in reducing overdrainage (10). Considerable controversy also revolves around the most optimum site for placement for the anti-siphon devices (66,67). The classic position is at the level of the skull base: however, the bench test suggests a marked tendency to overdrain if the SCD is below the level of the proximal catheter. These factors may be minor when considered in light of the excessive sensitivity of the SCD to external pressure from scar or when the patient is lying on the device (64).

The gravity actuated device (GAD) is used in conjunction with a differential pressure valve to limit overdrainage (68). It is similar to the horizontal vertical valve used in lumbar shunts but constructed to fit in-line with a ventriculoperitoneal shunt. There is no literature to prove or disprove its utility; however, in individual cases, we have found it effective. Experimental evidence suggests that motion and vibration (35) make the mechanism of these devices ineffective although clinical studies are lacking. The position of the GAD device is critical for optimum functioning. Slight angulation of the device to vertical can cause underdrainage in the horizontal position and overdrainage in the vertical position. Examples of pressure flow characteristics of a standard differential pressure valve, a flow control valve, and a valve containing a GAD are shown in Fig. 15.

Distal Catheter

Distal shunt malfunction is reported to occur in 12% to 34% of shunts (51,69). Three types of distal catheters have been used: the closed ended with side slits, open ended with side slits, and open ended. A higher incidence of distal catheter obstruction has been noted in catheters with side slits whether closed ended or open ended (51,70). Omental ingrowth is responsible for the peritoneal catheter obstruction; possibly the distal slits act as collection points for the debris and provide a channel for trapping the omentum. It is unclear whether using open-ended distal catheters increases the likelihood of small ventricle malfunction. Use of extended length catheters (110–120 cm) is not associated with an increase in the complications and eliminates the need to lengthen the peritoneal catheter for growth of the patient (71). However, care must be taken to identify patients who may have enough length of tubing in the abdomen but may underdrain due to a narrow and taught segment of tubing from subcutaneous tethering as a result of scarring and calcification.

It is difficulty to justify use of atrial over the peritoneal site for distal absorption (72,73). Data on 887 patients suggested that atrial shunts have a higher rate of malfunction although some studies have not shown a significant difference. However, when the same information was stratified by age, shunt type, and time period, there was no significant difference in shunt durability. Cardio-pulmonary complication, such as irreversible pulmonary hypertension, endocarditis, and glomerulonephritis, are some of the more serious complications that may occur with atrial shunts (73). Alternative sites, like pleura, may result in significant negative pressures in the shunt system (74). Poor absorption from the pleura may result in large pleural effusions in small children (74). The gall bladder has also been effectively used in patients in whom peritoneal, atrial, or pleural sites have been exhausted (75,76). Potential complications of these shunts, notably biliary ventriculitis and biliary meningitis, have been reported in the literature (77,78). The ventriculo-femoral shunt may be tried in patients with a difficult access to the atrium from the subclavian or jugular route (79). Trans-diaphragmatic placement of the distal catheter in the sub-hepatic space worked successfully in one reported patient with poor peritoneal access due to scarring (80).

Shunt Material

Ideal shunt material should be completely biocompatible, be easy to handle, flexible, resistant to infection, and nonmetallic but radio-opaque (metals interfere with MRI imaging). From a manufacturing standpoint, it should be easy

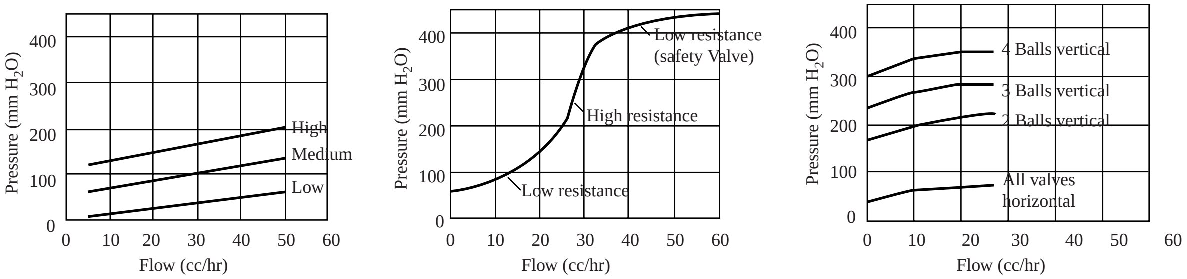


Figure 15. Pressure flow characteristics of (a). Standard differential pressure valve; note that with increasing differential pressure, such as from upright posture, there is an increase in the flow rate. (b) The flow control valve has a sigmoid flow-pressure relationship; in the upright position, the valve works at the high resistance stage and maintains a relatively steady flow rate despite increase in differential pressure. (c) Gravity actuated device, in vertical position acts as a very high-resistance differential pressure valve (depending on the number of balls in the device) and as a low-resistance differential pressure valve in the supine position.

to mold into tubing and making valve components. Silicone polymer is probably the best available material for this purpose.

Some studies have suggested development of silicone allergy in some patients with ventricular shunts (81–84). It is unclear whether it represents a true immunological reaction or a nonspecific foreign body type granulomatous reaction (85). In patients with suspected or documented silicone allergy, use of polyurethane (86) or, more recently CO_2 extracted silicone catheters has been postulated but not proven to offer some advantage in reducing risk of recurrent malfunctions.

Subcutaneous location of the distal catheters makes them susceptible to degradation from a foreign-body reaction mounted by the body (87). Scarring around the catheter, calcification, and stress fractures are long-term consequences of this reaction (88,89). Unless there is some amount of surface degradation, the adhesions to the subcutaneous tissues do not occur (87). Evidence suggests that barium used in the silicone catheters is probably not an important factor in promoting calcification and degradation (90). Use of barium-free catheters, however, makes it difficult to evaluate a shunt system on radiological imaging.

To minimize colonization of shunt catheters and infection, recently antibiotic-coated catheters have become available. The catheters are available coated with rifampin and minocycline (Medtronics, Goleta, CA) and another with rifampin and clindamycin (Bactiseal, Codman, Raynham, MA). The antibiotic is most active against Staph epidermidis, which is the cause of shunt infection in most patients. The antibiotic gradually leaches out of the catheter over a 30–60 day period providing added advantage. Control studies have shown a significant reduction in rate of infection with use of these catheters (91,92). The major drawback is the excessive cost of the antibiotic-coated catheters.

Shunt Malfunction

About 30–40% of the shunts malfunction within the first year of placement (10) and 80% of malfunctions are proximal malfunctions. Although most patients with a malfunctioning shunt will present with the classic features of raised pressure, headache, and vomiting, in 20%, there may be no signs of raised pressure (93). Instead, this group of patients present with a subtle change in behavior, decline in school performance, gait disturbances, and incontinence. Some patients may present with aggravation in the signs and symptoms of Chiari malformation or syringomyelia. Parents are often more sensitive to these subtle changes. In a study comparing the accuracy of referral source in diagnosing shunt malfunction, parents were more likely to be correct about the diagnosis as compared with a hospital or general practitioner (94).

At examination, a tense fontanelle, split sutures, and swelling at the shunt site are very strongly suggestive of a malfunctioning shunt. Shunt pumping has a positive predictive value of only 20% (95). A shunt valve that fails to fill up in 10 minutes is very strongly suggestive of shunt malfunction. Radiological assessment may demonstrate a fracture or dislocation. Presence of double-backing of the distal catheter, wherein the distal catheter tip loops out of the peritoneal through the same spot that it enters it, is diagnostic of distal malfunction (96). The shunt tap gives useful information about the proximal and distal shunt system. The absence of spontaneous flow and poor drip rate indicate proximal malfunction, whereas a high opening pressure is suggestive of distal malfunction (97). The presence of increase in size of the ventricles on CT scan confirms a malfunctioning shunt; however, a large number of patients with long-standing shunt have altered brain compliance and may not dilate the ventricles at the time of presentation. In children, similar symptoms occur in the common illnesses like otitis media; gastroenteritis of viral fevers often confound the diagnosis. Radiological assessment of shunt flow using radionuclide or iodide contrast media injected into the shunt may help (2,98–100). Unfortunately, although some studies have shown an accuracy of 99% with combined pressure and radionuclide evaluation (98,101), others have shown a 25–40% incidence of deceptive patency when evaluated by radionuclide cisternogram (97). This could stem from a partial but inadequately functioning shunt, intermittent malfunction, or presence of isolated ventricle. Similar problems are encountered with an iodide contrast-based shuntogram or shunt injection tests. In the absence of normative data with regard to adequate flow in the shunt, which may vary significantly with the individual, time of the day, and activity (102), use of Doppler-based flow devices, flow systems that work based on differential temperature-gradient or MRI-based flow systems becomes irrelevant for an individual patient. Lumbar infusion tests and shunt infusion tests to assess the outflow resistance through the shunt are cumbersome and require a laboratory-based setup and may not be possible in an ER setting (103–105). Infusion through a reservoir to assess outflow resistance through the shunt suggests a cutoff of less than 12 mm Hg/mL/min as reliable for distinguishing a clinically suspected high probability of malfunction from those with a low probability of shunt malfunction (104). However, this is the group of patients who may not really need the test, and patients who have a questionable malfunction on clinical grounds often have equivocal results on the infusion study.

In childhood hydrocephalus, ICP is the only accurate guide to shunt function other than the symptoms (105). Again the ability to measure ICP through the valve tap becomes unreliable with a partial proximal malfunction. A similar problem may be encountered with in-line telemetric ICP monitors (106,107). In-addition, the telemetric transducers may develop a significant drift over time. In difficult cases, the only way to resolve the issue may be to explore the shunt, to measure ICP through a lumbar puncture if the patient has communicating hydrocephalus, or to place an ICP monitor. Noninvasive monitoring of ICP is going to have a major role in assessment of these patients. For patients who have a very compliant brain, ventricular diltation on the CT scan easily confirms inadequate shunt function.

Despite advances in shunt technology, the incidence of shunt malfunction has not changed over the last 50 years. Nulsen and Becker (3) reported a rate of malfunction of 44%, in 1967, which is similar to that reported in recent studies. To improve on the existing shunt systems, it is

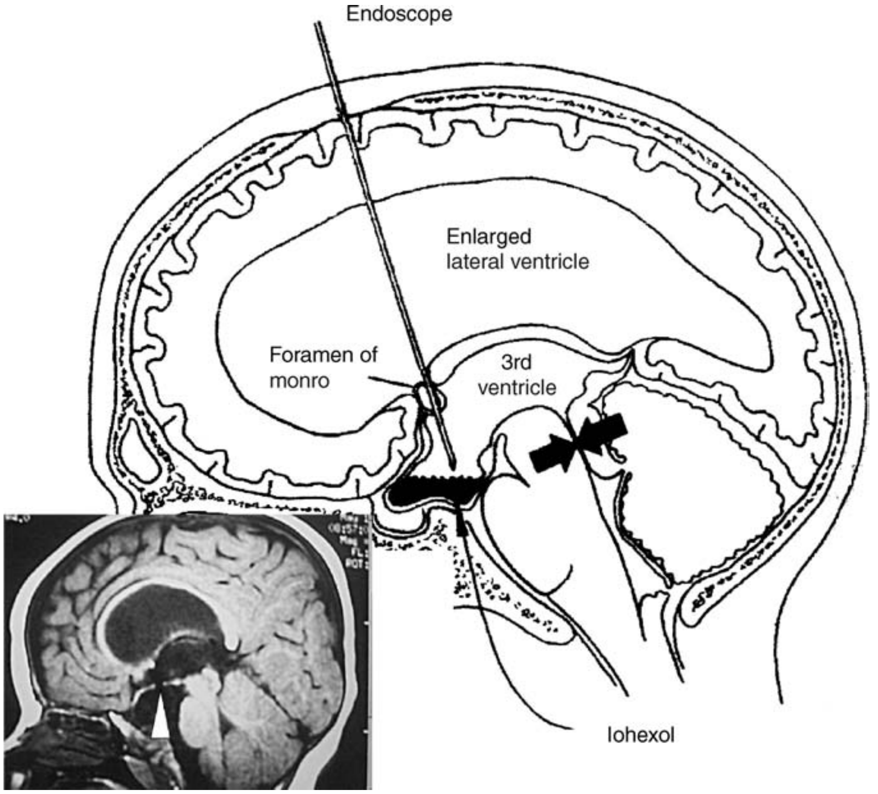


Figure 16. Third ventriculostomy is useful in patients with obstructive hydrocephalus such as observed with aqueductal stenosis (arrows). An endoscope is passed through a small hole in the skull into the frontal horn of the ventricle and navigated into the third ventricle. The floor of the third ventricle is then perforated under vision (arrow, in the inset) so as to bypass the obstruction at the aqueduct. Iohexol, a nonionic iodinated dye, is then instilled into the ventricle to demonstrate good communication between the ventricles and the subarachnoid space.

important to understand the factors that reduce shunt survival. Although the location of the proximal catheter has not been clearly shown to influence shunt survival (50,51), the presence of a small amount of fluid around the proximal catheter is associated with longer shunt survival. In a study that looked at shunt over a 11 year period, statistically significant differences were noted in shunt survival in patients with tumor versus post-hemorrhagic and aqueductal stenosis; shunts in infants and the pediatric age group survive shorter than in adults; shunt after multiple revisions survive shorter, and additional shunts placed for isolated ventricles have shorter survival (70,108). Chronic inflammatory changes of granular ependymitis often seen at the time of endoscopic shunt placement in patients with multiple revisions probably contribute to recurrent malfunction and progressive shortening of the interval between revisions as the number of surgeries increase (108).

The nature of the valve clearly influences the risk of proximal catheter malfunction. It is much lower with flow control valves (10,109). Overdrainage from the differential pressure valves pulls the choroid plexus toward the proximal catheter and may promote malfunction (55). However, the increased rate of valve malfunction in flow control devices balances out this advantage (10).

THIRD VENTRICULOSTOMY

In patients with an obstructive type of hydrocephalus, third ventriculostomy offers an alternative to shunt. The procedure involves making an opening in the relatively thin membrane of the floor of the third ventricle. This is accomplished by passing an endoscope through the lateral ventricle and guiding it through the foramen of Munro to the floor of the third ventricle (Fig. 16). The opening allows CSF to bypass the obstruction at the level of the aqueduct or the fourth ventricle and directly enter the subarachnoid space.

Although third ventriculostomy has been recommended as a procedure of choice for obstructive hydrocephalus; data from some prospective studies have failed to show an improved cure rate (110,111). A retrospective analysis of ventriculographic versus endoscopic third ventriculostomy in 213 cases does show the superiority of the endoscopic procedure over the ventriculographic operation both in terms of reduced risk and improved survival of the procedure (112). Despite the theoretical advantages, evidence suggests that third ventriculostomy may not be effective in controlling raised intracranial pressure in all patients (112,113). Early failures in a radiologically proven case of obstructive hydrocephalus may relate to multifactorial etiology of hydrocephalus; associated absorption defects, obliteration of subarachnoid space from long-standing ventricular diltation, and unidentified infectious cause of aqeductal stenosis may be responsible. Late failures may relate to gliotic scarring over the ventriculostomy, which has been visually confirmed by endoscope in some cases. Does the ventriculostomy close from scarring, or is it a secondary response to lack of flow through the ventriculostomy due to poor absorption, therefore, a lack of gradient between the ventricle and the subarachnoid space, is unclear. In a small prospective study comparing the shunt failure rate with the failure rate of third ventriculostomy, no statistical difference was found between the two (12). Likewise, no controlled study has compared laser, blunt, or sharp fenestration of the floor or demonstrated usefulness of balloon diltation of the fenestration. The success rate of third ventriculostomy of 49–100%, as reported in the literature, may not be a true representative of the efficacy of third ventriculostomy (12). Evaluation after third ventriculostomy and defining success is difficult in the absence of documented reduction in ICP or improvement on neuropyschological tests. This is so because the ventricles may not reduce in size, and to and from motion through the patent fenestration may still be observed on MRI CSF flow studies even though the patient may be symptomatic. In the absence of clear evidence in literature, third ventriculostomy is often advocated for patients with obstructive hydrocephalus, and if failures occur, shunting is preferred over repeat fenestration.

It is hoped that close collaboration between the industry and medicine will help develop "smart shunts" that would be able to mimic physiological CSF dynamics. These devices will possibly incorporate nanotechnology and would be superior to the presently available devices. It is likely that better understanding of CSF drainage mechanism in the future may help develop alternatives such as drugs that improve drainage of CSF through lymphatic/arachnoidal venous channels or promote proliferation of new lymphatic/arachnoidal venous channels. Until that time, it seems that shunts are the best available alternative for management of communicating hydrocephalus.

BIBLIOGRAPHY

1. Dandy WE, Blackfan KD. Internal hydrocephalus: An experimental, clinical and pathological study. Am J Dis Child 1914;30:406–482.
2. Graham P, Howman-Giles R, Johnston I, Besser M. Evaluation of CSF shunt patency by means of technetium-99m DTPA. J Neurosurg 1982;57:262–266.
3. Nulsen FE, Becker DP. Control of hydrocephalus by valve-regulated shunt. J Neurosurg 1967;26:362–374.
4. Pudenz RH, Russell FE, Hurd AH, Shelden CH. Ventriculo-auriculostomy; a technique for shunting cerebrospinal fluid into the right auricle; preliminary report. J Neurosurg 1957;14:171–179.
5. Nulsen FE, Spitz EB. Treatment of hydrocephalus by direct shunt from ventricle to jugular vain. Surg Forum 1951;94: 399–403.
6. Fernell E, Hagberg B, Hagberg G, von Wendt L. Epidemiology of infantile hydrocephalus in Sweden. I. Birth prevalence and general data. Acta Paediatr Scand 1986;75:975–981.
7. Bondurant CP, Jimenez DF. Epidemiology of cerebrospinal fluid shunting. Pediatr Neurosurg 1995;23:254–258; discussion 259.
8. Foltz EL, Shurtleff DB. Five-year comparative study of hydrocephalus in children with and without operation (113 Cases). J Neurosurg 1963;20:1064–1079.
9. Laurence KM, Coates S. The natural history of hydrocephalus. Detailed analysis of 182 unoperated cases. Arch Dis Child 1962;37:345–362.
10. Drake JM, Kestle JR, Milner R, Cinalli G, Boop F, Piatt J, Jr. Haines S, Schiff SJ, Cochrane DD, Steinbok P, MacNeil N. Randomized trial of cerebrospinal fluid shunt valve design in pediatric hydrocephalus. Neurosurgery 1998;43:294–303; discussion 303–295.
11. Patwardhan RV, Nanda A. Implanted ventricular shunts in the United States: The billion-dollar-a-year cost of hydrocephalus treatment. Neurosurgery 2005;56:139–144; discussion 144–135.
12. Tuli S, Alshail E, Drake J. Third ventriculostomy versus cerebrospinal fluid shunt as a first procedure in pediatric hydrocephalus. Pediatr Neurosurg 1999;30:11–15.
13. Bell WO. Cerebrospinal fluid reabsorption. A critical appraisal. 1990. Pediatr Neurosurg 1995;23:42–53.
14. Edsbagge M, Tisell M, Jacobsson L, Wikkelso C. Spinal CSF absorption in healthy individuals. Am J Physiol Regul Integr Comp Physiol 2004;287:R1450–1455.
15. Koh L, Zakharov A, Johnston MG. Integration of the subarachnoid space and lymphatics: Is it time to embrace a new concept of cerebrospinal fluid absorption? Cerebrospinal Fluid Res 2005;2:6.
16. Johnston M, Zakharov A, Papaiconomou C, Salmasi G, Armstrong D. Evidence of connections between cerebrospinal fluid and nasal lymphatic vessels in humans, non-human primates and other mammalian species. Cerebrospinal Fluid Res 2004;1:2.
17. Greitz D, Greitz T, Hindmarsh T. We need a new understanding of the reabsorption of cerebrospinal fluid—II. Acta Paediatr 1997;86:1148.
18. Marmarou A, Shulman K, LaMorgese J. Compartmental analysis of compliance and outflow resistance of the cerebrospinal fluid system. J Neurosurg 1975;43:523–534.
19. Raksin PB, Alperin N, Sivaramakrishnan A, Surapaneni S, Lichtor T. Noninvasive intracranial compliance and pressure based on dynamic magnetic resonance imaging of blood flow and cerebrospinal fluid flow: Review of principles, implementation, and other noninvasive approaches. Neurosurg Focus 2003;14:e4.
20. McAllister JP2nd, Chovan P. Neonatal hydrocephalus. Mechanisms and consequences. Neurosurg Clin N Am 1998; 9:73–93.
21. Adams RD, Fisher CM, Hakim S, Ojemann RG, Sweet WH. Symptomatic occult hydrocephalus with "normal" cerebrospinal-fluid pressure. A treatable syndrome. N Engl J Med 1965; 273:117–126.
22. Oi S, Shimoda M, Shibata M, Honda Y, Togo K, Shinoda M, Tsugane R, Sato O. Pathophysiology of long-standing overt ventriculomegaly in adults. J Neurosurg 2000;92:933–940.
23. Di Rocco C, Pettorossi VE, Caldarelli M, Mancinelli R, Velardi F. Experimental hydrocephalus following mechanical increment of intraventricular pulse pressure. Experientia 1977;33:1470–1472.
24. Egnor M, Rosiello A, Zheng L. A model of intracranial pulsations. Pediatr Neurosurg 2001;35:284–298.
25. Egnor M, Zheng L, Rosiello A, Gutman F, Davis R. A model of pulsations in communicating hydrocephalus. Pediatr Neurosurg 2002;36:281–303.
26. Linninger AA, Tsakiris C, Zhu DC, Xenos M, Roycewicz P, Danziger Z, Penn R. Pulsatile cerebrospinal fluid dynamics in the human brain. IEEE Trans Biomed Eng 2005;52:557–565.
27. Stephensen H, Tisell M, Wikkelso C. There is no transmantle pressure gradient in communicating or noncommunicating hydrocephalus. Neurosurgery 2002;50:763–771; discussion 771–763.
28. Portnoy HD, Branch C, Castro ME. The relationship of intracranial venous pressure to hydrocephalus. Childs Nerv Syst 1994;10:29–35.
29. Marmarou A. A theoretical model and experimental evaluation of the cerebrospinal fluid system. 1973.
30. Alperin N, Lichtor T, Mazda M, Lee SH. From cerebrospinal fluid pulsation to noninvasive intracranial compliance and pressure measured by MRI flow studies. Curr Med Imaging Rev. In press.

31. Marmarou A, Shulman K, Rosende RM. A nonlinear analysis of the cerebrospinal fluid system and intracranial pressure dynamics. J Neurosurg 1978;48:332–344.
32. Ekstedt J. CSF hydrodynamic studies in man. 1. Method of constant pressure CSF infusion. J Neurol Neurosurg Psych 1977;40:105–119.
33. Lundkvist B, Eklund A, Kristensen B, Fagerlund M, Koskinen LO, Malm J. Cerebrospinal fluid hydrodynamics after placement of a shunt with an antisiphon device: A long-term study. J Neurosurg 2001;94:750–756.
34. Czosnyka M, Batorski L, Laniewski P, Maksymowicz W, Koszewski W, Zaworski W. A computer system for the identification of the cerebrospinal compensatory model. Acta Neurochirurgica 1990;105:112–116.
35. Borgesen SE, Gjerris F. The predictive value of conductance to outflow of CSF in normal pressure hydrocephalus. J Neurol 1982;105:65–86.
36. Marmarou A, Bergsneider M, Klinge P, Relkin N, Black PM. The value of supplemental prognostic tests for the preoperative assessment of idiopathic normal-pressure hydrocephalus. Neurosurgery 2005;57:17–28.
37. Huang TY, Chung HW, Chen MY, Giiang LH, Chin SC, Lee CS, Chen CY, Liu YJ. Supratentorial cerebrospinal fluid production rate in healthy adults: Quantification with two-dimensional cine phase-contrast MR imaging with high temporal and spatial resolution. Radiology 2004;233:603–608.
38. Alperin NJ, Lee SH, Loth F, Raksin PB, Lichtor T. MR-Intracranial pressure (ICP): A method to measure intracranial elastance and pressure noninvasively by means of MR imaging: Baboon and human study. Radiology 2000;217:877–885.
39. Urchuk SN, Plewes DB. MR measurements of pulsatile pressure gradients. J Magn Reson Imaging 1994;4:829–836.
40. Bird R, Stewart W, Lightfoot E. Transport Phenomena. New York: Wiley Sons; 1960.
41. Walters BC, Hoffman HJ, Hendrick EB, Humphreys RP. Cerebrospinal fluid shunt infection. Influences on initial management and subsequent outcome. J Neurosurg 1984;60: 1014–1021.
42. Pople IK, Ettles D. The role of endoscopic choroid plexus coagulation in the management of hydrocephalus. Neurosurgery 1995;36:698–701; discussion 701–692.
43. Weiss MH, Nulsen FE, Kaufman B. Selective radionecrosis of the choroid plexus for control of experimental hydrocephalus. J Neurosurg 1972;36:270–275.
44. Aoki N. Lumboperitoneal shunt: Clinical applications, complications, and comparison with ventriculoperitoneal shunt. Neurosurgery 1990;26:998–1003; discussion 1003–1004.
45. Selman WR, Spetzler RF, Wilson CB, Grollmus JW. Percutaneous lumboperitoneal shunt: Review of 130 cases. Neurosurgery 1980;6:255–257.
46. Chumas PD, Armstrong DC, Drake JM, Kulkarni AV, Hoffman HJ, Humphreys RP, Rutka JT, Hendrick EB. Tonsillar herniation: The rule rather than the exception after lumboperitoneal shunting in the pediatric population. J Neurosurg 1993;78:568–573.
47. Payner TD, Prenger E, Berger TS, Crone KR. Acquired Chiari malformations: Incidence, diagnosis, and management. Neurosurgery 1994;34:429–434; discussion 434.
48. Ausman JI. Shunts: Which one, and why? Surg Neurol 1998;49:8–13.
49. Albright AL, Haines SJ, Taylor FH. Function of parietal and frontal shunts in childhood hydrocephalus. J Neurosurg 1988;69:883–886.
50. Bierbrauer KS, Storrs BB, McLone DG, Tomita T, Dauser R. A prospective, randomized study of shunt function and infections as a function of shunt placement. Pediatr Neurosurg 1990;16:287–291.
51. Sainte-Rose C, Piatt JH, Renier D, Pierre-Kahn A, Hirsch JF, Hoffman HJ, Humphreys RP, Hendrick EB. Mechanical complications in shunts. Pediatr Neurosurg 1991;17:2–9.
52. Pang D, Grabb PA. Accurate placement of coronal ventricular catheter using stereotactic coordinate-guided free-hand passage. Technical note. J Neurosurg 1994;80:750–755.
53. Yamamoto M, Oka K, Nagasaka S, Tomonaga M. Ventriculoscope-guided ventriculoperitoneal shunt and shunt revision. Technical note. Acta Neurochir (Wien) 1994;129:85–88.
54. Kestle JR, Drake JM, Cochrane DD, Milner R, Walker ML, Abbott R, 3rd, Boop FA. Lack of benefit of endoscopic ventriculoperitoneal shunt insertion: A multicenter randomized trial. J Neurosurg 2003;98:284–290.
55. Hakim S. Observations on the physiopathology of the CSF pulse and prevention of ventricular catheter obstruction in valve shunts. Dev Med Child Neurol Suppl 1969;20:42–48.
56. Martinez-Lage JF, Lopez F, Poza M, Hernandez M. Prevention of intraventricular hemorrhage during CSF shunt revisions by means of a flexible coagulating electrode. A preliminary report. Childs Nerv Syst 1998;14:203–206.
57. Steinbok P, Cochrane DD. Removal of adherent ventricular catheter. Pediatr Neurosurg 1992;18:167–168.
58. Whitfield PC, Guazzo EP, Pickard JD. Safe removal of retained ventricular catheters using intraluminal choroid plexus coagulation. Technical note. J Neurosurg 1995;83: 1101–1102.
59. Handler MH. A complication in removing a retained ventricular catheter using electrocautery. Pediatr Neurosurg 1996; 25:276.
60. Czosnyka M, Czosnyka Z, Whitehouse H, Pickard JD. Hydrodynamic properties of hydrocephalus shunts: United Kingdom Shunt Evaluation Laboratory. J Neurol Neurosurg Psych 1997;62:43–50.
61. Trost HA. Is there a reasonable differential indication for different hydrocephalus shunt systems? Childs Nerv Syst 1995;11:189–192.
62. Sood S, Kumar CR, Jamous M, Schuhmann MU, Ham SD, Canady AI. Pathophysiological changes in cerebrovascular distensibility in patients undergoing chronic shunt therapy. J Neurosurg 2004;100:447–453.
63. Pollack IF, Albright AL, Adelson PD. A randomized, controlled study of a programmable shunt valve versus a conventional valve for patients with hydrocephalus. Hakim-Medos Investigator Group. Neurosurgery 1999;45:1399–1408; discussion 1408–1311.
64. Aschoff A, Kremer P, Benesch C, Fruh K, Klank A, Kunze S. Overdrainage and shunt technology. A critical comparison of programmable, hydrostatic and variable-resistance valves and flow-reducing devices. Childs Nerv Syst 1995;11:193–202.
65. Drake JM, da Silva MC, Rutka JT. Functional obstruction of an antisiphon device by raised tissue capsule pressure. Neurosurgery 1993;32:137–139.
66. Fox JL, Portnoy HD, Shulte RR. Cerebrospinal fluid shunts: An experimental evaluation of flow rates and pressure values in the anti-siphon valve. Surg Neurol 1973;1:299–302.
67. Tokoro K, Chiba Y. Optimum position for an anti-siphon device in a cerebrospinal fluid shunt system. Neurosurgery 1991;29:519–525.
68. Chabbra DK, Agarwal GD, Mittal P. "Z" flow hydrocephalus shunts, a new approach to the problem of hysrocephalus. The rationale behind its design and the initial results of pressure monitoring after "Z" flow shunt implantation. Acta Neurochir (Wien) 1993;121:43–47.
69. Sekhar LN, Moossy J, Guthkelch AN. Malfunctioning ventriculoperitoneal shunts. Clinical and pathological features. J Neurosurg 1982;56:411–416.
70. Cozzens JW, Chandler JP. Increased risk of distal ventriculoperitoneal shunt obstruction associated with slit valves or

distal slits in the peritoneal catheter. J Neurosurg 1997;87: 682–686.

71. Couldwell WT, LeMay DR, McComb JG. Experience with use of extended length peritoneal shunt catheters. J Neurosurg 1996;85:425–427.

72. Borgbjerg BM, Gjerris F, Albeck MJ, Hauerberg J, Borgesen SV. A comparison between ventriculo-peritoneal and ventriculo-atrial cerebrospinal fluid shunts in relation to rate of revision and durability. Acta Neurochir (Wien) 1998;140: 459–464; discussion 465.

73. Lam CH, Villemure JG. Comparison between ventriculoatrial and ventriculoperitoneal shunting in the adult population. Br J Neurosurg 1997;11:43–48.

74. Willison CD, Kopitnik TA, Gustafson R, Kaufman HH. Ventriculopleural shunting used as a temporary diversion. Acta Neurochir (Wien) 1992;115:67–68.

75. Ketoff JA, Klein RL, Maukkassa KF. Ventricular cholecystic shunts in children. J Pediatr Surg 1997;32:181–183.

76. Novelli PM, Reigel DH. A closer look at the ventriculo-gallbladder shunt for the treatment of hydrocephalus. Pediatr Neurosurg 1997;26:197–199.

77. Barami K, Sood S, Ham SD, Canady AI. Postural changes in intracranial pressure in chronically shunted patients. Pediatr Neurosurg 2000;33:64–69.

78. Bernstein RA, Hsueh W. Ventriculocholecystic shunt. A mortality report. Surg Neurol 1985;23:31–37.

79. Philips MF, Schwartz SB, Soutter AD, Sutton LN. Ventriculofemoroatrial shunt: A viable alternative for the treatment of hydrocephalus. Technical note. J Neurosurg 1997;86: 1063–1066.

80. Rengachary SS. Transdiaphragmatic ventriculoperitoneal shunting: Technical case report. Neurosurgery 1997;41: 695–697; discussion 697–698.

81. Goldblum RM, Pelley RP, O'Donell AA, Pyron D, Heggers JP. Antibodies to silicone elastomers and reactions to ventriculoperitoneal shunts. Lancet 1992;340:510–513.

82. Gower DJ, Lewis JC, Kelly DL, Jr. Sterile shunt malfunction. A scanning electron microscopic perspective. J Neurosurg 1984;61:1079–1084.

83. Snow RB, Kossovsky N. Hypersensitivity reaction associated with sterile ventriculoperitoneal shunt malfunction. Surg Neurol 1989;31:209–214.

84. Sugar O, Bailey OT. Subcutaneous reaction to silicone in ventriculoperitoneal shunts. Long-term results. J Neurosurg 1974;41:367–371.

85. Kalousdian S, Karlan MS, Williams MA. Silicone elastomer cerebrospinal fluid shunt systems. Council on Scientific Affairs, American Medical Association. Neurosurgery 1998; 42:887–892.

86. Jimenez DF, Keating R, Goodrich JT. Silicone allergy in ventriculoperitoneal shunts. Childs Nerv Syst 1994;10: 59–63.

87. Del Bigio MR. Biological reactions to cerebrospinal fluid shunt devices: A review of the cellular pathology. Neurosurgery 1998;42:319–325; discussion 325–316.

88. Echizenya K, Satoh M, Murai H, Ueno H, Abe H, Komai T. Mineralization and biodegradation of CSF shunting systems. J Neurosurg 1987;67 :584–591.

89. Elisevich K, Mattar AG, Cheeseman F. Biodegradation of distal shunt catheters. Pediatr Neurosurg 1994;21:71–76.

90. Irving IM, Castilla P, Hall EG, Rickham PP: Tissue reaction to pure and impregnated silastic. J Pediatr Surg 1971;6:724–729.

91. Aryan HE, Meltzer HS, Park MS, Bennett RL, Jandial R, Levy ML. Initial experience with antibiotic-impregnated silicone catheters for shunting of cerebrospinal fluid in children. Childs Nerv Syst 2005;21:56–61.

92. Govender ST, Nathoo N, van Dellen JR: Evaluation of an antibiotic-impregnated shunt system for the treatment of hydrocephalus. J Neurosurg 2003;99:831–839.

93. Fried A, Shapiro K. Subtle deterioration in shunted childhood hydrocephalus. A biomechanical and clinical profile. J Neurosurg 1986;65:211–216.

94. Watkins L, Hayward R, Andar U, Harkness W. The diagnosis of blocked cerebrospinal fluid shunts: A prospective study of referral to a paediatric neurosurgical unit. Childs Nerv Syst 1994;10:87–90.

95. Piatt JH, Jr. Physical examination of patients with cerebrospinal fluid shunts: Is there useful information in pumping the shunt? Pediatrics 1992;89:470–473.

96. Martinez-Lage JF, Poza M, Izura V. Retrograde migration of the abdominal catheter as a complication of ventriculoperitoneal shunts: The fishhook sign. Childs Nerv Syst 1993;9: 425–427.

97. Sood S, Canady AI, Ham SD. Evaluation of shunt malfunction using shunt site reservoir. Pediatr Neurosurg 2000; 32:180–186.

98. Hayden PW, Rudd TG, Shurtleff DB. Combined pressure-radionuclide evaluation of suspected cerebrospinal fluid shunt malfunction: A seven-year clinical experience. Pediatrics 1980;66:679–684.

99. Sweeney LE, Thomas PS. Contrast examination of cerebrospinal fluid shunt malfunction in infancy and childhood. Pediatr Radiol 1987;17:177–183.

100. Vernet O, Farmer JP, Lambert R, Montes JL. Radionuclide shuntogram: Adjunct to manage hydrocephalic patients. J Nucl Med 1996;37:406–410.

101. Savoiardo M, Solero CL, Passerini A, Migliavacca F. Determination of cerebrospinal fluid shunt function with water-soluble contrast medium. J Neurosurg 1978;49:398–407.

102. Kadowaki C, Hara M, Numoto M, Takeuchi K, Saito I. CSF shunt physics: factors influencing inshunt CSF flow. Childs Nerv Syst 1995;11:203–206.

103. Czosnyka M, Whitehouse H, Smielewski P, Simac S, Pickard JD. Testing of cerebrospinal compensatory reserve in shunted and non-shunted patients: A guide to interpretation based on an observational study. J Neurol Neurosurg Psych 1996;60:549–558.

104. Morgan MK, Johnston IH, Spittaler PJ. A ventricular infusion technique for the evaluation of treated and untreated hydrocephalus. Neurosurgery 1991;29:832–836; discussion 836–837.

105. Fouyas IP, Casey AT, Thompson D, Harkness WF, Hayward RD. Use of intracranial pressure monitoring in the management of childhood hydrocephalus and shunt-related problems. Neurosurgery 1996;38:726–731; discussion 731–722.

106. Woodford J, Saunders RL, Sachs E, Jr. Shunt system patency testing by lumbar infusion. J Neurosurg 1976;45:60–65.

107. Cosman ER, Zervas NT, Chapman PH, Cosman BJ, Arnold MA. A telemetric pressure sensor for ventricular shunt systems. Surg Neurol 1979;11:287–294.

108. Miyake H, Ohta T, Kajimoto Y, Matsukawa M. A new ventriculoperitoneal shunt with a telemetric intracranial pressure sensor: Clinical experience in 94 patients with hydrocephalus. Neurosurgery 1997;40:931–935.

109. Lazareff JA, Peacock W, Holly L, Ver Halen J, Wong A, Olmstead C. Multiple shunt failures: An analysis of relevant factors. Childs Nerv Syst 1998;14:271–275.

110. Decq P, Barat JL, Duplessis E, Leguerinel C, Gendrault P, Keravel Y. Shunt failure in adult hydrocephalus: Flow-controlled shunt versus differential pressure shunts–a cooperative study in 289 patients. Surg Neurol 1995;43:333–339.

111. Garton HJ, Kestle JR, Cochrane DD, Steinbok P. A cost-effectiveness analysis of endoscopic third ventriculostomy. Neurosurgery 2002;51:69–77; discussion 77–68.

112. Santamarta D, Diaz Alvarez A, Goncalves JM, Hernandez J. Outcome of endoscopic third ventriculostomy. Results from an unselected series with noncommunicating hydrocephalus. Acta Neurochir (Wien) 2005;147:377–382.
113. Cinalli G, Sainte-Rose C, Chumas P, Zerah M, Brunelle F, Lot G, Pierre-Kahn A, Renier D. Failure of third ventriculostomy in the treatment of aqueductal stenosis in children. J Neurosurg 1999;90:448–454.
114. Hirsch JF, Hirsch E, Sainte Rose C, Renier D, Pierre-Khan A. Stenosis of the aqueduct of Sylvius. Etiology and treatment. J Neurosurg Sci 1986;30:29–39.

See also Intrauterine surgical techniques; microdialysis sampling; monitoring, intracranial pressure.

HYPERALIMENTATION. See Nutrition, parenteral.

HYPERBARIC MEDICINE

Barbara L. Persons
Beth Collins
William C. Lineaweaver
University of Mississippi
Medical Center
Jackson, Mississippi

INTRODUCTION

The goal of hyperbaric oxygen (HBO) therapy is to deliver high concentrations of oxygen under pressure to increase the amount of dissolved oxygen in the blood. The physiologic repercussions of this increased plasma oxygen have widespread effects that translate into a variety of clinical applications. Initially, the use of hyperbaric medicine surrounded acute decompression illness and gas embolism. Later, the increased oxygen under pressure was shown to have use in a variety of clinical situations. The delivered pressure can be two to six times ambient atmospheric pressure (ATM) or atmospheres absolute (ATA) depending on the indication. The current Undersea and Hyperbaric Medical Society approved uses of HBO are shown in Table 1, and many of these indications will be discussed individually. HBO in the treatment of these conditions is supported by controlled medical trials published in peer-reviewed journals and, as such, is evidence-based. Numerous other experimental uses exist for HBO, such as for stroke and for cardiac ischemia, but these uses have not yet been sufficiently proven to be supported by the Undersea and Hyperbaric Medicine Society or by the American College of Hyperbaric Medicine. The goal of this chapter is to outline the physical principles underlying the use of HBO therapy, to discuss its medical indications for HBO, and to familiarize the reader with the mechanical, safety, and regulatory issues involved in operating a hyperbaric medicine program.

Table 1. Approved Uses for Hyperbaric Oxygen Therapy

acute decompression illness
gas embolism
carbon monoxide poisioning
clostridial gas gangrene
necrotizing soft tissue infections
compromised skin grafts and skin flaps
crush injury
compartment syndrome
acute traumatic ischemias
radiation tissue damage
refractory osteomyelitis
selected problem wounds
acute exceptional blood loss anemia
acute thermal burns
intracranial abscess

HISTORICAL BACKGROUND

British physician and clergyman Henshaw was the first to use alteration in atmospheric pressure to treat medical conditions when he used his domicilium chamber in 1862. Hyperbaric medicine also surrounded diving and diving medicine. Triger, in 1841, gave the first human description of decompression sickness (1). In 1934, U.S. Naval Submarine Officer Dr. Albert Behnke was the first to use oxygen recompression to treat decompression sickness in naval divers (2). Later, in 1943, Gagnon and Cousteau invented SCUBA (self-contained underwater breathing apparatus). Dr. Boerma, a Dutch thoracic surgeon, removed the blood cells from pigs in 1955 and found they could survive with the oxygen dissolved in plasma by use of HBO. An upsurge in hyperbaric surgery followed in 1956, when Boerma performed cardiovascular surgery in a hyperbaric chamber, which along with hypothermia, allowed for periods of circulatory arrest of 7–8 min. The large chamber developed at Wilhelmina Gasthuis in Amsterdam in 1959, headed by Boerma, allowed a wide variety of research to be carried out on the uses of HBO therapy on many diseases (3). By 1966, it was indicated for the treatment of protection during induced circulatory arrest, homotransplantation, clostridial infection, acute arterial insufficiency, chronic arterial insufficiency, and hypovolemic shock. Shortly thereafter, the advent of cardiopulmonary bypass obviated hyperbaric chambers for cardiac protecton. In 1967, the Undersea Medical Society was founded by the U.S. Navy diving and submarine medical officers. This organization originally focused on undersea and diving medicine but, later, came to include clinical hyperbaric medicine. In 1986, the name was changed to the Undersea and Hyperbaric Medical Society or UHMS with more than 2500 physician and scientist members in 50 countries. More recently, the American College of Hyperbaric Medicine has come to offer board certification to U.S. physicians in the specialty of hyperbaric medicine.

PHYSICS

To understand HBO therapy, one must understand a few basic laws of physics, namely, Boyle's law, Charles law, Dalton's law, and Henry's law. Boyles law explains how gas volume shrinks with increasing pressure. Charles law explains that the volume of a gas decreases with decreasing temperature. Dalton's law explains that each gas in a mixture exerts its own partial pressure independently of the others. Henry's law explains that the number of gas molecules that will dissolve in a liquid depends on the partial pressure of the gas as well as on the mass of the liquid. These laws themselves will be explained in the first part of this section and the application of the laws to each area of hyperbaric medicine will be explained in the respective section.

Boyle's Law

Boyles law states if the temperature remains constant, the volume of a gas is inversely proportional to the pressure.

Boyle's law is stated as $PV = K$ or $P = K/V$, where P is pressure, V is volume, and K is a constant.

Thus, as in Table 2, the volume of a bubble at 1 Atmosphere or sea level shrinks to one-half of its original volume at 33 feet below sea level (10 m or 2 atmospheres). It shrinks to one-third of its original size at 66 feet below sea level (20 m or 3 atmospheres). Conversely, if a diver were to hold his breath at depth and then ascend, the air in his lungs will expand to three times the volume it occupied at 66 feet below sea level.

Charle's Law

Charles' law states that if the pressure remains constant, the volume of a fixed mass of gas is directly proportional to absolute temperature. The volume increases as the temperature increases. For example, a balloon has a volume of 1 L at 20 °C and its volume would expand to 1.1 L at 50 °C.

Charles' law is stated as $V/T = K$ or $V1/T1 = V2/T2$, where V is volume of gas 1 or 2 , T is temperature in Kelvin of gas 1 or 2, and K is a constant.

Dalton's Law

Dalton's law states that each gas in a mixture exerts its partial pressure independently, which is important in understanding human physiogy and HBO. For example, if a patient is given a high concentration of oxygen and a lower concentration of nitrogen, these gasses will diffuse across membranes and act in the body independently of each other.

Dalton's law is stated as $P(t) = P1{+}P2{+}P3$, where $P(t)$ is total pressure and $P1$, $P2$, and $P3$ are the individual gas pressures. Gases try to equalize their concentrations across a membrane, which explains how breathing a higher oxygen, lower nitrogen mixture can help nitrogen leave the blood through the lungs as nitrogen gas.

Henry's Law

Henry's law states that at a given temperature, the amount of gas dissolved in a solute is directly proportional to the pressure of the gas above the substance.

Table 2. Pressure vs. Volume at Depth

Depth (feet sea water)	Pressure (ATA) atmospheres	Gas Volume (%)
0(sea level)	1	100
33	2	50
66	3	33
165	6	17

Henry's law illustrates that when a liquid is exposed to a gas, some of the gas molecules dissolve into it. The number of moles that will dissolve in the liquid depends on the mass of the liquid, the partial pressure of the gas, its solubility in the liquid, the surface area of contact, and the temperature (as that changes with the partial pressure). Thus, more gas, oxygen, or nitrogen will dissolve in tissue fluid at a higher pressure because the partial pressure of each gas increases at a higher pressure.

Henry's law is stated as $p = Kc$, where p is the partial pressure of the gas 1, c is its molar concentration, and K is the Henry's law constant, which is temperature-dependent.

An example of Henry's law is dissolved carbon dioxide in soda, which bubbles out of solution as the pressure decreases. Another example is exemplified by water when it is heated. Long before it boils, bubbles of air form on the side of the pan, which is an example of the gas coming out of solution as the temperature is raised.

For treatment of decompression illness (DCI) and gas embolism (GE), increased pressure alone as well as the increased oxygen pressure facilitate treatment. Both of these conditions are a result of gas bubbles in the tissues or gas bubbles in the blood causing blockage of vessles or ischemia of tissues. Therefore, shrinking the bubbles with increased pressure allows them to be removed or minimized by the body, which is the principle of Boyle's law, that the volume of a gas varies inversely with pressure. If one has a bubble occluding an important vessel or lodged in a joint, it will shrink as the pressure increases as in Table 2. The blood can accommodate an increased amount of dissolved gas with increased atmospheric pressure as explained by Henry's law. Henry's law states that the amount of gas that will dissolve in a liquid is proportional to the partial pressure of the gas in contact with that liquid as well as to the atmospheric pressure. Atmospheric pressure at sea level is 760 mmHg, and the normal atmosphere consists of 21% oxygen and 79% nitrogen. (see Fig. 1) (4). The pressure of a gas dissolved in plasma relates to its solubility in a liquid as well as to its partial pressure. A gas bubble caught in the tissues or in the systemic circulation either because of an air embolus or because of decompression sickness is significantly decreased under hyperbaric conditions, as illustrated by Table 3. By Boyles law, the pressure alone shrinks the bubble to a fraction of its' original size. Dalton's law states that total pressure of a mixture of gases is equal to the sum of the pressures of each gas. Each gas is acting as if it alone were present, which explains how the oxygen under pressure creates a situation whereby the higher dissolved oxygen surrounds the bubble and causes the diffusion of nitrogen out of the bubble, called nitrogen washout. Simply put, each gas attempts to have equal concentration of particles, in this case, nitrogen and oxygen on each side of the gas bubble. Nitrogen, therefore, diffuses out of the bubble and shrinks in size. Hyperbaric oxygen enables treatment of the two conditions where air or nitrogen bubbles become lodged in the tissues.

PHYSIOLOGY

Hyperbaric oxygen therapy oxygen enters the systemic circulation through the alveoli in the lungs, and the

Table 3. Oxygen Levels During Hyperbaric Oxygen Treatment Breathing Air (4)

ATA Chamber	Pressure Chamber	PO_2 (mmHg) Chamber	PAO_2 (mmHg) Lung	O_2 ml/dl vol % Plasma
1	760	160	100	0.31
2	1520	319	269	0.83
2.36	1794	377	322	1.00
2.82	2143	450	400	1.24
3	2280	479	429	1.33
4	3040	638	588	1.82
5	3800	798	748	2.32
6	4560	958	908	2.81
Breathing 100% Oxygen				
1	760	760	673	2.08
2	1520	1520	1433	4.44
2.36	1794	1794	1707	5.29
2.82	2143	2143	2056	5.80
3	2280	2280	2193	6.80
4	3040	100% oxygen is not used above 3ATA to minimize the risk of oxygen toxicity.		
5	3800			
6	4560			

diffusion is mediated by the pressure differential between the alveolar oxygen content and the oxygen content of venous blood. Alveolar oxygen content is 100 mmHg and venous oxygen content is 40 mm Hg (5). Normally, 97% of oxygen iscarried in the arterial blood bound to hemoglobin molecules and only 3% of the oxygen is dissolved in plasma, illustrated by the formula for oxygen content in arterial blood. $CaO_2 = (1.34 \times \mathrm{Hb} \times \mathrm{SaO_2}) + (0.003 \times PaO_2)$, where CaO_2 is oxygen content, Hb is Hemoglobin in grams, and SaO_2 is arterial O_2 saturation expressed as a fraction not a percentage (0.95, not 95%). 1.34 is the realistic binding capacity of hemoglobin, although 1.39 is the actual binding capacity. 0.003 times the PaO_2 is the amount of oxygen soluble in plasma at normal atmospheric pressure. With hyperbaric oxygen, the amount of oxygen dissolved in arterial blood is dramatically increased. Conversely, the amount carried by hemoglobin remains about the same as that achieved by inspiring oxygen at 1 atmosphere absolute (ATA) (4). As measured in ml O_2 per deciliter of whole blood, the oxygen content increases significantly under hyperbaric conditions as shown in Table 3. The oxygen content of blood increases from 0.31 at 1 atmosphere to 6.80ml/dl vol% at 3 atmospheres. Note that 100% oxygen is not used at pressures greater than 3 ATA to minimize the risk of oxygen toxicity. The alveolar type 1 cells in the lungs and the neurons in the brain are sensitive to excessive concentrations of oxygen. Again, to prevent oxygen toxicity, which can lead to alveolar damage or seizure, 100% oxygen is not given at pressures greater than 3 ATA. Even in these pressure ranges, air breaks are given to patients and they breathe air as opposed to concentrated oxygen under pressure for 10 minutes during many of the protocols, which in theory, gives the xanthine oxidese system a chance to deal with the current load of free radicals in the lungs and tissues and the Gaba amino buteric acid (GABA) depletion in the brain a chance to normalize. As mentioned, the blood's ability to carry more dissolved oxygen molecules under higher atmospheric pressure is a function of Henry's law. It states that the amount of gas that will dissolve in a liquid is proportional to the partial pressure of the gas in contact with that liquid as well as to the atmospheric pressure. Thus, more gas, oxygen, or nitrogen will dissolve in tissue fluid at higher atmospheric pressure. %X is the percentage of gas X dissolved in the liquid, $P(t)$ is the atmospheric pressure, and $P(\mathrm{X})$ is the partial pressure of gas X.

TRANSCUTANEOUS OXYMETRY (TcPO$_2$ OR TCOM)

In order for the patient to benefit from HBO therapy, they must have adequate perfusion of blood to the affected area. This perfusion can be assessed prior to hyperbaric oxygen therapy by checking transcutaneous oxygen tension, $TcPO_2$. Transcutaneous oxygen tension values of less than 40, which increase to more than 100 mmHg while breathing 100% oxygen or to more than 200 during HBO therapy, will likely benefit from hyperbaric oxygen therapy (6). The detailed mechanisms by which the elevated oxygen tension is felt to improve wound healing and the body's ability to combat bacterial pathogens will be discussed under each indication for HBO therapy. Briefly, Hunt and Pa: (7) showed, in 1976, that increased oxygen tension stimulated collagen synthesis and fiberblast proliferation. Then, studies revealed improved ability of leukocytes to clear infected wounds of bacteria with hyperbaric oxygen (8,9). Then, in 1989, Zamboni et al. (10) showed that HBO therapy in the reduction of tissue flap necrosis was a systemic phenomenon that involved inhibition of neutrophil adherence and prevention of arteriolar vasoconstriction thought to be via a nitric oxide-mediated mechanism. In addition, it increases platelet-derived growth factor Beta (PDGF B), vascular endothelial growth factor (VEGF), Epidermal growth factor (EGF), and other factors.

APPROVED INDICATIONS

The following indications are approved uses of hyperbaric oxygen therapy as defined by the Hyperbaric Oxygen

Therapy Committee of the Undersea and Hyperbaric Medical Society (Table 1). Most of these indications are also covered by Medicare and many are covered by major insurance companies. The indications include air or gas embolism, decompression illness, carbon monoxide poisioning, Clostridial myositis and myonecrosis (gas gangrene), crush injury, compartment syndrome and other acute traumatic ischemias, decompression sickness, problem wounds, exceptional blood loss anemia, intracranial abscess, necrotizing soft tissue infections, refractory osteomyelitis, delayed radiation injury, compromised skin grafts and flaps, and thermal burns. Many of these indications will be specifically discussed in the following sections.

HYPERBARIC CHAMBER BASICS

Hyperbaric oxygen therapy is a feature offered in many hospitals, medical centers, and in specialty situations such as diver rescue stations and oil rigs. Over 500 hyperbaric chambers exist in the United States alone. When a patient is refered for one of the listed emergent or nonemergent indications to undergo hyperbaric oxygen therapy, a complete workup of the patient should be performed if possible prior to hyperbaric oxygen therapy. Emergency situations may necessitate an abbreviated exam. The patient should wear only cotton medical-center-provided clothing to prevent static electricity, which could cause a spark and a fire. All foreign appliances should be removed, including hearing aids, lighters, and jewelry. Internal appliances such as pacemakers are usually safe under hyperbaric conditions. The patient will then be premedicated with a benzodiazapine such as valium if needed for anxiety. Hyperbaric oxygen therapy can be administered through monoplace or multiplace chambers. Monoplace chambers accommodate a single patient within a pressurized environment of 100% oxygen (Fig. 1). These chambers are often constructed as acrylic cylinders with steel ends and can withstand pressures of up to 3 ATA. Multiplace chambers are usually constructed of steel and can withstand pressures up to 6 ATA (Fig. 2). They can accommodate two or more people and often have the capacity to treat ventilated or critically ill patients. Some are even large enough to accommodate operating teams, and the 100% oxygen is delivered to the patient via face mask, hood, or endotrachial tube. Depending on the indication, patients will require from 1 to 60 treatments. Most commonly, treatments are delivered to the patient for 60–90 min at 2.8–3.0 ATA five days a week for the protocol duration.

CONTRAINDICATIONS

Six absolute contraindications exist to hyperbaric oxygen therapy.

1. Untreated pneumothorax – An untreated pneumothorax can be converted to a tension pneumothorax with administration of HBO.
2. History of spontaneous pneumothorax.

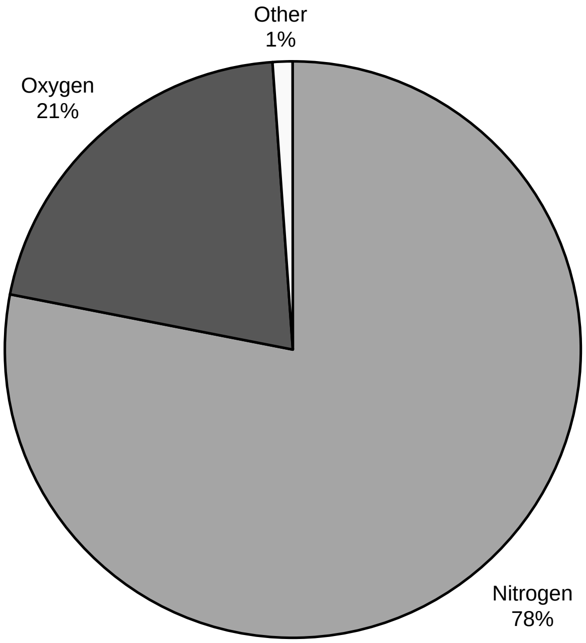


Figure 1. Relative composition of air.

3. Bleomycin – History of the chemotherapy agent Bleomycin, which can cause pneumonitis, especially if the patient is exposed to HBO.
4. Doxyrubicin.
5. Disulfiram – (antibuse) Blocks production of superoxide dismutase, which protects the patient from oxygen toxicity.
6. Cisplatin/Carboplatin – An anticancer agent that interferes with DNA synthesis.
7. Mefanide (Sulfamyelon) – It is a topical ointment for burns and wounds that is a carbonic anhydrase inhibitor and increases the risk of seizure during HBO therapy.

The relative considerations and contraindications are many. In these patients, the hyperbaric physician should consider each patient individually, including their history of thoracic surgery, seizure disorder, obstructive lung disease, congestive heart failure, pulmonary lesions on X-ray or CT scan, seizure disorder, upper respiratory infections

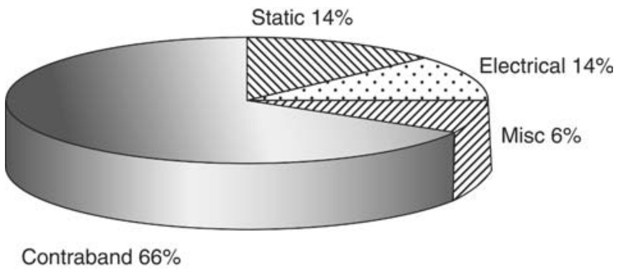


Figure 2. Fire risk.

Table 4. Arterial and Venous Air Embolism Etiology

Etiology of Air Embolism	
Arterial Air Embolism	*Venous Air Embolism*
Pulmonary Overpressure	Central Venous Catheters
Open Lung Biobsy	Infusion Pumps
Arterial Catheter	Neurosurgery
Angiography	Laparoscopy
Surgery	Liver Transplantation
Penetrating Chest Trauma	Neurosurgery
Pneumothorax	Pelvic Surgery Trendelenburg Position
Cardiopulmonary Bypass	Necrotizing Enterocolitis
Dialysis	Lumbar Spine Spine Surgery
Autotransfusion	Air Contrast Salpingogram
Neonatal Respiratory Distress Syndrome	Umbilical Venous Catheters
Paradoxical, Patent Foramen Ovale	

(due to risk of barotraumas to the ears), acute viral infections, uncontrolled high fever (as it increases CNS sensitivity to oxygen), reconstructive ear surgery, congenital spherocytosis, history of optic neuritis, recent retinal repair, claustrophobia, acidosis, nicotine, alcohol, and many others.

AIR EMBOLISM

Air embolism is a medical emergency. In the diving community, it is the main cause of death following diving accidents. Early diagnosis followed by definitive treatment are critical in determining the eventual outcome. Treatment is based on compression of the air bubbles by Boyle's law as well as oxygenation of ischemic tissues and treatment of ischemia reperfusion injury with hyperbaric oxygen. Air embolism can occur in the hospital setting by introduction of air into the systemic circulation by central venous and arterial catheters and other invasive procedures. Interestingly, the first report of a death from air embolism was from France in 1821. The patient was undergoing surgery on his clavicle when the surgeon noted a hissing sound in the area of operation and the patient yelled "my blood is falling into my heart- I'm dead" (11). The patient likely died of a venous air embolism obstructing the systemic circulation. Air entry in the systemic circulation occurs following violation of the systemic circulation by any number of mechanisms (Table 4), which can be either by introduction of air into the arterial circulation, as in a lung biopsy, chest trauma, and pulmonary overpressure (diving), or into venous circulation, as in air introduction via central venous catheters, liver transplantation, and neurosurgery. Venous air emboli are more common, whereas arterial emboli are tend to be more serious. Physiologically, the air bubble forms or is introduced into the circulation. The lung usually serves as an excellent filter for air emboli, and can protect the embolism from traveling to the brain. This protective filter may be bypassed by a patent foramen ovale. Approximately 30% of patients have a foramen ovale that is patent by probe. The lung as a filter may be overwhelmed by large quantities of air. The bubble can then lodge in the smaller arteries of the brain causing obstruction. An air embolism is immediately identified as a foreign body, and platelets are activated, which leads to an inflammatory cascade. Hypoxia then develops distal to the obstruction with associated swelling. The embolism is eventually absorbed by the body, but the fibrin deposition at the embolism site may prevent return of blood flow. In order to diagnose an air embolism, it needs to have a high index of suspicion. If air embolism is suspected, the patient should receive a number of immediate measures, including ACLS or ATLS protocols. The patient should be placed on the left side and one should consider draining air from the right atrium with a central venous catheter. Air in the heart can cause a machinery murmur. 100% oxygen should be administered via a face mask or endotrachial tube, and the patient should be hydrated to preserve intravascular volume. Per Dalton's law, administration of high oxygen will cause nitrogen to diffuse out of the air bubble and shrink in size. Dalton's law states that total pressure of a mixture of gases is equal to the sum of the pressures of each gas. Each gas acts as if it alone were present. Dalton's law is stated as $P(t) = P1 + P2 + P3$. The inspired 100% oxygen will also maximize oxygenation of the tissues as much as possible under normal atmospheric pressure. Next, the patient should be emergently treated with hyperbaric oxygen. If a chamber is available that can provide compression to 6 ATA with air or a mixture of 50% nitrogen and 50% air, treatment should immediately be performed following the U.S. Navy protocol (Fig. 3).

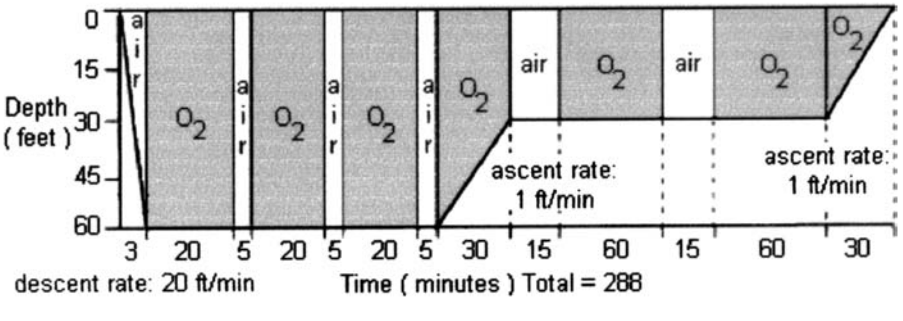


Figure 3. U.S. Navy decompression treatment.

DECOMPRESSION ILLNESS

By definition, decompression illness (DCI), also called bends or caisson disease, occurs when gas bubbles exit in the blood or body tissues. At depth, more gas can dissolve in the tissues than in the blood. When the diver ascends too quickly, the gas comes out of solution and forms bubbles in the tissues and in the blood, much like popping a soda can and releasing its pressure causes the carbon dioxide bubbles to come out of the solution. In 1994, Diver's Alert Network (DAN) recorded 1164 diving-related injuries and 97 diving-related deaths, many related to DCI (12). The severity of DCI depends on the volume and location of gas bubbles. The range of symptoms is from vague constitutional complaints or limb pain to cardiopulmonary arrest and coma, the pathophysiology of which is explained by Henry's law. As previously explained, Henry's law is $p = Kc$, where p is the partial pressure of the gas 1, c is its molar concentration, and K is the Henry's law constant, which is temperature-dependent. Thus, simply stated, more inert gas can be dissolved in a liquid at higher atmospheric pressure and, conversely, less under lower atmospheric pressure, which occurs on decompression when gas must be removed from tissues, and rapid decompression leads to bubble formation. Using the Henry's law equation, one can calculate the estimated amount of nitrogen a diver must clear from the bloodstream (about 5 L) in rising from 100 ft to the surface. The amout would be approximately 750 ml nitrogen assuming room temperature, which is a significant volume of nitrogen that must be eliminated from the divers bloodstream. The onset of symptoms is usually rapid and 75% of patients experience symptoms within 1 h of decompression and 90% within 12 h of decompression (13). A small number of patients may present even later, particularly if they have flown in commercial aircraft after diving and not followed the recommendation of the major diving organizations not to fly within 24 hours of one's last dive. Interestingly, up to 10% of the inert gas that is absorbed in the tissues is released as bubbles after the diver's decompression (14). Patients experience symptoms depending on the location and concentration of the bubbles (see Table 5). Bubbles forming in or near joints cause the joint pain of a classical "bend." These musculoskeletal effects are called type 1 DCS. When these effects occur in the spinal cord or brain, numbness, paralysis, and disorders of higher cerebral function may result. If large numbers of bubbles enter the venous bloodstream, congestive symptoms in the lung and circulatory shock can then occur. These pulmonary and neurologic effects are termed type 2 DCS. Treatment should involve immediate administration of 100% oxygen, which facilitates nitrogen washout by the previously explained principles of Dalton's law. Rehydration as well as advanced cardiac or trauma life-support protocols should be followed by transfer to a hyperbaric facility emergently. The patient should be treated with hyperbaric oxygen following U.S. Navy guidelines (Fig. 3), even if the inspired oxygen and rehydration alone have improved the patient's signs and symptoms because tiny bubbles may be left that can cause tissue necrosis. Hyperbaric oxygen shrinks the size of the mostly nitrogen-filled bubbles by the principles of Boyle's law, and the increased pressure also increases the partial pressure of the gas by Dalton's law, hastening complete elimination of the bubble (Table 2). If the U.S. Navy (Fig. 3) recompression regimen fails to lead to symptom resolution, the Diver's Alert Network or a medical expert on DCI should be contacted, and one of a number of recompression tables may be followed.

Table 5. Signs and Symptoms of Decompression Illness

Symptoms of Decompression Illness	Signs of Decompression Illness
Unusual fatigue	Blotchy skin rash
Skin itching	Paralysis, muscle weakness
Pain in joints/ muscles of arms, legs or torso	Difficulty urinating
Dizziness	Confusion, personality changes, bizarre behavior
Vertigo	Amnesia,
Ringing in the ears, (Tinnitus)	Staggering
Numbness, tingling and paralysis	Coughing up bloody, frothy sputum
Shortness of breath (Dyspnea)	Collapse or unconsciousness
	Tremors

PROBLEM WOUNDS

The management of problem wounds should always include infection control, debridement, aggressive wound care, and correction of perfusion and oxygenation deficiencies. When an oxygenation deficiency of the wound is found, in the face of nonreconstructable vascular disease, hyperbaric oxygen should be considered as an adjunctive therapy. An increase in tissue oxygen tension by HBO therapy enhances wound healing by increasing neutrophil bactericidal capacity, inhibiting toxin formation in and even killing some anaerobes, encouraging fibroblast activity, and promoting angiogenesis (15). In normal physiology, the oxygen gradient across a wound is essential to stimulate these components of healing. Oxygen consumption is relatively low in wounds, and microvasculature damage and peripheral vasoconstriction increase diffusion distances. Partial pressure via Dalton's law is the driving force of diffusion. Hyperbaric oxygen creates a steep tissue oxygenation gradient, providing a stronger stimulus than lactate or moderate hypoxia, to initiate and facilitate wound healing (16,17). These stimulated factors are thought to include platelet-derived growth factor B (PDGF-B), Vascular endothelial growth factor (VEGF), Epidermal growth factor, and others. Several clinical studies support the use of hyperbaric oxygen to promote wound healing. Perhaps the studies involving diabetic lower extremity wounds have been most informative. Several studies have shown an increased number of healed wounds, decreased wound size, and decreased rates of amputation among patients receiving hyperbaric oxygen therapy as an adjunctive treatment (18,19). Baroni et al. reported in a controlled study that a significant number of subjects receiving HBO went on to heal their wounds and fewer required amputation when compared with subjects not receiving HBO (20). In another study involving 151 diabetic patients with wounds of the lower extremity, Oriani et al. showed that 130 of these patients completely healed their wounds with adjunctive HBO (21). When compared with conventionally treated wounds, HBO-treated patients had an accelerated rate of healing, reduced rate of amputation, and an increased rate of completely healed wounds on a long-term basis (21). Transcutaneous Oxymetry (TcPO_2) is currently the best tool available to evaluate tissue hypoxia and to select patients appropriate for HBO therapy. It can also be used to monitor progress during hyperbaric oxygen therapy. TcPO_2 measurements should be taken with the patient breathing room air. A value of greater than 50 mmHg around the wound site indicates that the wound has adequate oxygenation and hyperbaric oxygen is not likely to improve healing. Values below 40 at the wound site should be considered for HBO therapy. Patients with marginal TcPO_2 should be further tested while breathing 100% oxygen. TcPO_2 values of greater than 100 while breathing 100% oxygen is an indicator that they are likely to respond to HBO therapy. If this challenge TcPO_2 is less than 100, they still may benefit if the tested TcPO_2 at the wound site is greater than 200 mmHg while they are breathing 100% oxygen at 2.0 ATA in the hyperbaric chamber (22). A TcPO_2 value of less than 30 around a wound that does not exhibit this response, which indicates vascular compromise and the patient should be considered for revascularization if possible. Of note, 96% of limbs with TcPO_2 values below 30 mmHg had abnormal arteriograms. It is also important to follow TcPO_2 values weekly, and diabetic patients may have normal or falsely elevated noninvasive Doppler studies and a low TcPO_2, implying satisfactory perfusion and inadequate oxygenation of the wound and, as such, may pose a diagnostic delimma. The diabetic patient with normal noninvasive Doppler and low TcPO_2 will respond best to HBO. HBO therapy should be reserved for those diabetic wounds not responding to traditional management of debridement, antibiotics, and general wound care, including vascular reconstruction. The use of HBO therapy is necessary in only 15–20% of these patients. HBO therapy increases wound oxygen tension, enhancing host antibacterial mechanisms and promoting wound healing and is reserved for wounds in which the primary etiologies are tissue hypoxia or infection (13). Treatments are delivered at 2.0–2.4 atmospheres for 90–120 min once or twice daily. When serious infections are present, patients are typically hospitalized and given IV antibiotics and hyperbaric treatments twice daily five days a week. The TcPO_2 values should be checked weekly because hyperbaric oxygen facilitates angiogenesis by a nitric oxide and vascular endothelial growth factor Beta (VEGF-B). When the room air TcPO_2 is greater than 40 mmHg, the hyperbaric oxygen therapy can safely be discontinued. HBO is an adjuvant treatment; therefore, diabetic control, debridement, and aggressive wound treatment are given first priority. When the wound bed has adequate granulation tissue, application of grafts can shorten morbidity, hospital stay, and health-care costs. The underlying problem in failure of a wound to heal is usually hypoxia and infection. Hyperbaric oxygen treatments in selected patients can facilitate healing by increasing tissue oxygen tension, thus providing the wound with a more favorable environment for healing. Therefore, hyperbaric oxygen therapy can be an important component to any comprehensive wound care program.

COMPROMISED FLAPS AND GRAFTS

Skin grafts and flaps with adequate blood supply do not require HBO. Hyperbaric oxygen therapy is extremely useful in situations where the skin grafts or flaps suffer from compromised microcirculation or hypoxia.

Flaps

The benefits of HBO on flaps develop from a systemic elevation in oxygen tension (23–25). In addition, HBO therapy prevents neutrophil adherence and subsequent vasoconstriction following ischemia. Too often, a compromised flap is allowed to progress over the days following surgery until visible signs of necrosis obviate the use of HBO, because delayed treatment with HBO cannot revive dead tissue. The resulting disappointment, as well as the associated patient dissatisfaction, can be avoided by rapid diagnosis of the flap problem and early involvement of the hyperbaric physician. The keys to successful treatment of compromised flaps with HBO are accurate diagnosis of the specific flap problem and appropriate and expedient initiation of hyperbaric oxygen treatment. Awareness of the different etiologies of flap compromise is necessary to plan for effective HBO treatment. A random flap with distal necrosis is completely different from a free flap with total venous occlusion. Proper classification of flaps, different etiologies of flap compromise, and understanding of how HBO is thought to effect ischemia reperfusion injury defines which patients will benefit from HBO. Flap classification is based on an assessment of blood supply, tissue composition, and method of movement. Each of these elements must be evaluated, but it is blood supply that is most important. The blood supply to the flap is either axial, based on a named vessel, or random, based on the subdermal plexus. Commonly, flap compromise occurs when the surgeon tries to mobilize tissue outside the defined arterial supply, when there is a pedicle problem exists, or when free flaps are exposed to prolonged ischemia. The tissue composition of a flap may include skin, subcutaneous tissue, fascia, muscle, bone, other tissues, or a combination of these. Flap composition is very important because different tissue types have different tolerances to ischemia. For instance, a myocutaneous flap will be more susceptible to ischemia than a fasciocutaneous flap, because muscle is much more sensitive to ischemic injury than fascia and skin (26). In those circumstances where a prolonged primary ischemia or any secondary ischemia resulting from vessel thrombosis and revision anastomosis exists, the flaps will undergo ischemia reperfusion injury.

When treating compromised flaps, a multimodality approach should be initiated. This approach should include the use of vasodilators if arterial vasospasm is suspected, removal of sutures if tension or compression are suspected, dextran and pentoxifylline for rheological purposes, medicinal and chemical leeching for venous congestion, and the early use of hyperbaric oxygen if blood flow can be documented. The use of HBO therapy is appropriate only when the flap problem has been defined, documented perfusion of the flap exists, appropriate surgical salvage measures have been first considered, and HBO therapy can be performed in an expedient manner. Specifically with respect to free flaps, extended primary ischemia time greater than 2 h or any secondary ischemia time may result in partial or total flap necrosis. This injury is usually reversible if recognized early and treated expeditiously. Essentially, it is ischemia reperfusion injury. Numerous research studies support the use of HBO in the salvage of compromised free tissue transfers (27,28). A rat free-flap model showed similar improvement in flap survival (27). A clinical study evaluated free-flap salvage in the face of prolonged primary or any secondary ischemia (28). Salvage was significantly better in the HBO treatment group vs. controls, but only if initiated within 24 h. Free flaps compromised by prolonged primary or secondary ischemia have responded favorably to HBO treatment with complete salvage, in most cases, if HBO is started early. The treatment regimen is 2.0–2.4 ATA, 90 min q 8 h x 24 h, then q 8–12 h x 48 h (29). Treatment duration is based on clinical evaluation.

Grafts

Skin grafts are anatomically different from flaps in that skin grafts lack an inherent blood supply. Skin grafts are composed of avascular tissue that depends entirely on the recipient bed for oxygenation. HBO is useful in preparing the recipient bed and in promoting healthy granulation tissue to support split-thickness skin grafts. One controlled study showed a significant improvement in skin graft survival from 17% to 64% with the addition of HBO treatment. Although literature exists to support the use of HBO for composite grafts, a study by the University of Mississippi Medical Center found no significant effect of HBO on rat-ear composite grafts larger than 1 cm (30,31) Further research is needed to better understand the effects of HBO on composite graft survival. The rational for use of HBO in crush injury, compartment syndrome, frostbite, and other traumatic ischemias is similar to those for compromised flaps as they are all cases of ischemia and ischemia reperfusion injury.

CRUSH INJURY, COMPARTMENT SYNDROME, AND OTHER ACUTE TRAUMATIC ISCHEMIAS

These conditions are trauma-related situations in which the underlying pathophysiology is that of ischemia reperfusion (IR) injury. Ischemia times of greater than 4 h willresult in some degree of permanent necrosis. The physiologic basis of IR injury has become better understood in recent years. Most of the animal research centers around the production of oxygen-free radicals. Although the endothelial xanthine oxidase pathway has received much attention in the literature (32), more recent evidence supports the fact that neutrophils are a more important source of oxygen-free radicals via membrane NADPH oxidase and degranulation. Also, neutrophil adhesion is felt to cause ischemia reperfusion IR-associated vasoconstriction.

A perceived paradox exists related to HBO for IR injury. The less-informed observer often does not understand why HBO improves reperfusion injury and might think HBO instead increases free radical formation. (An oxygen-free radical is an oxygen molecules with an unpaired electron in its outer shell.) During ischemia, ATP is ultimately degraded to hypoxanthine and xanthine, which are anaerobic metabolites. With reperfusion, oxygenated blood is reintroduced into the ischemic tissue, and the hypoxanthine and xanthine plus oxygen creates oxygen-free radicals. Superoxide and hydroxyl oxygen-free radicals are formed, which can cause extensive tissue damage. The authors believe that the major mediator of damage is, in fact, neutrophil adherence to postcapillary venules significant and progressive vasoconstriction occurs in arterioles adjacent to leukocyte-damaged venules. Neutrophil adherence and vasoconstriction lead to a low flow state in the microcirculation and then vessel thrombosis, which is the endpoint of IR injury. The leukocyte-damaged venule is thought to be responsible for the arterial vasoactive response. HBO inhibits neutrophil adherence to the endothelial cells and thereby inhibits the ultimate thrombosis of microvessels, but the complete mechanism is still poorly understood, but is thought to involve the elevation in nitric oxide mediated by an increase in nitric oxide syntase (33). Free radical formation is not felt to be worsened with HBO as fewer adherent neutrophils actually exist to contribute to the neutrophil oxygen-free radical-generating system.

Treatment with hyperbaric oxygen in the face of IR injury carried the concern that that providing extra oxygen would increase free radical production and tissue damage. This query has been resolved by studies that have shown that HBO actually antagonizes the ill effects of IR injury in a variety of tissues (33–35). One of the first studies evaluating HBO and IR injury showed that HBO, immediately upon reperfusion, significantly improved skin flap survival following 8 h of global ischemia in a rat axial skin flap model with increased microvascular blood flow during reperfusion. Free-flap survival improves with HBO treatment during reperfusion even following ischemia times of up to 24 h (36). Hyperbaric oxygen administered during and up to 1 h following 4 h global ischemia significantly reduced neutrophil endothelial adherence in venules and also blocked the progressive arteriolar vasoconstriction associated with reperfusion injury (37). HBO inhibited *in vitro* beta-2-integrin (CD18)-induced neutrophil adherence function, but did not alter other important neutrophil functions such as oxidative burst or stimulus-induced chemotaxis and migration. This latter finding is very important, because HBO, through its action on the CD18 adhesion molecule, blocks the neutrophil adherence associated with IR injury without interfering with other neutrophil functions that would increase the risk of infectious complications. Initially, the focus in acute ischemia caused by trauma should be restoration of blood supply. The authors, therefore, recommend HBO therapy for all patients with muscle ischemia time greater than 4 h and skin ischemia time greater than 8 h. The major effects of IR injury are felt to occur within the first 4–7 h of reperfusion. 2 ATA hyperbaric oxygen increases the tissue oxygen tension 1000%. Treatment protocol is 2.0–2.5 ATA for 60 min, q 8 h x 24 h, then q 8–12 h x 48 h with clinical re-evaluation. If progressive signs of ischemic injury are still present, the treatment is continued at 2.0 ATA, q 12 h for 2–3 more days. Usually, 72 h of treatment is adequate as long as the first treatment is initiated within 4 h of surgery.

RADIATION TISSUE DAMAGE AND OSTEORADIONECROSIS

1.2 million cases of invasive cancer are diagnosed yearly, half of which will receive radiation therapy and 5% of which will have serious radiation complications, which represents 30,000 cases per year of serious radiation sequellae (38). HBO is also well studied for its use in treating osteoradionecrosis in conjunction with adequate debridement of necrotic bone. Carl et al. also reported success is applying HBO to 32 women with radiation injury following lumpectomy and radiation compared with controls (39). Feldmeier and his colleagues reviewed the literature and found no evidence to support the potentiation of malignant cells or the engancement of cancer growth (40). The treatment protocol is 2.5 ATA for 90 min daily for 20–50 treatments. HBO can also be used as a radiosensitizer and are as much as three times more sensitive to radiation kisses than are hypoxic cells (41).

REFRACTORY OSTEOMYELITIS

Chronic refractory osteomyelitis (CROM) is infection of the medullary and cortical portions of the bone that persists or recurs following treatment with debridement and antibiotics. The principles of treatment are fairly simple. First, the dead bone is debrided and bone cultures should be taken along with administration of appropriate antibiotics. Next, the interface or cicatrix, which separates the compromised bone from adequate blood supply, is removed. Finally, hypoxia in the wound must be corrected, which may be accomplished by HBO. The treatment protocol is 2.0 ATA for 90 min daily for 20–60 treatments. Note that CROM and refractory osteomyelitis require the longest treatment protocols. HBO is believed to oxygenate hypoxic/ischemic tissues, augment host antimicrobial responses, augment osteoclastic activity, and induce osteogenesis in normal and infected bone and antibiotic synergism.

ACUTE THERMAL BURNS

HBO is approved by the USMS but it is not covered by Medicare. Gruber demonstrated in 1970 that the area around and under a third-degree burn was hypoxic and could only be raised by oxygen at increased pressure (42). HBO has been found to prevent extension, reduce edema, increase healing rates, and decrease total cost in several randomized studies (43,44). HBO is also thought to decrease the rate of burn sepsis based on several early studies. The controversy, in part, surrounds current guidelines for early debridement and grafting of burns. Once excised, a burn no longer exists and HBO will not be helpful. In case burns are not easily amenable to excision such as flash burns to the face or groin, HBO may be helpful to prevent extension of the burn and to aid healing. Treatment must be started within 24 h. The recommended regimen is 2.0 ATA for 90 min every 8 h on the first day, then every 12 hours for 5 or 6 days.

ACUTE EXCEPTIONAL BLOOD LOSS ANEMIA

Hyperbaric oxygen for treatment of acute blood loss anemia is reserved for those patients whose anemia is not immediately treatable for practical, disease process, or religious reasons, which may include warm antibody hemolytic disease, Jehova's Witnesses, those with rare blood types, and those who refuse transfusion for other personal reasons. As explained in the physiology section, HBO dramatically increases the amount of solublized oxygen the blood can carry. In fact, Boerema showed, in 1955, that pigs could be exsanguinated to four-tenths of one gram of hemoglobin per deciliter and be maintained in a hyperbaric environment of 3 ATA without hypoxia. The goal in HBO therapy for these conditions is to improve the oxygen depth with the daily or twice daily HBO treatments until the anemia can be improved. In between the treatments, the patients should be maintained a lower $F\mathrm{IO}_2$ of inspired oxygen if possible to help reduce oxygen toxicity.

CARBON MONOXIDE POISONING

In 1966, Wada first used HBO to treat survivors of coal mine disasters with carbon monoxide poisoning and burns. The modern-day sources of carbon monoxide include automobile exhaust, home heaters, portable generators, propane engines, charcoal burners and camp stoves, and methylene chloride paint strippers. The initial treatment for carbon monoxide poisoning is 100% oxygen. The administration of 100% oxygen via a nonrebreather mask facilitates the dissociation of CO from hemoglobin to approximately 1.5 h. Hyperbaric oxygen delivered at 2.8–3.0 ATA reduced the halflife of CO-bound hemoglobin further to 23 min. In addition, patients who had one hyperbaric treatment for CO poisoning had 46% neuropsychiatric sequelae at discharge and 50% at one month versus two HBO treatments at 2.8–3.0 ATA having 13% at discharge and 18% at one month. The current recommendation is 3.0 ATA for 90 min with air breaks delivered every 8 h for a total of 3 treatments (called theWeaver protocol). Some authors still feel one treatment may be adequate (45).

CYANIDE POISONING

Hydrocyanide gas or HCN is formed when any number of substances burns, including furniture, asphalt, paper, carpeting (nylon), lighting baths (acrylic), plastic(polystyrene), and insulation (melamine resins). The antidote for cyanide poisoning begins with breathing 100% oxygen, ATLS protocols, and administration of IV sodium thiosulphate and is continued with a slow infusion of sodium nitrate and simultaneous HBO therapy if it is available. The sodium nitrate creates methemoglobin, which can impair the oxygen-carrying capacity of hemoglobin. HBO increases the amount of oxygen dissolved in plasma and may offer a direct benefit. The treatment regimin is 3.0 ATA with 30/10 airbreaks.

HYPERBARIC CHAMBER FACILITY DESIGN AND SAFETY

Over 500 hyperbaric facilities exist in the United States, and the number of hyperbaric chambers is steadily increasing worldwide. Hyperbaric chambers are classified as either monoplace or multiplace. They differ functionally in that the monoplace chamber instills oxygen into the entire chamber environment, whereas in a multiplace chamber, patients breathe 100% oxygen via a breathing mask or oxygen hood and exhaled gases are vented outside the chamber. Monoplace chambers are constructed either as an acrylic cylinder with metal ends or are primarily constructed of metal. Most commonly, the monoplace chambers are formed from an acrylic cylinder from 20 to 40 inches in diameter with tie rods connecting it to end caps. The opening is a rotating lock or a cam action lever closure. Separate oxygen and air sources provide the oxygen sources and air for air breaks during therapy. An oxygen vent must be exhausted outside the building. The through ports on the HBO chamber door allow passage of specially made intravenous monitoring devices and ventilators. The larger diameter monoplace chambers are more comfortable; however, they require more oxygen and can be heavier and more expensive to install. The acrylic chambers can provide a maximum of 3 ATA pressure. Alternatively, some monoplace chambers are constructed mostly of steel with acrylic view ports, which can accommodate pressures of up to 6 ATA and are often used in special situations such as offshore rigs where a compact chamber is needed to treat decompression illness required in U.S. Navy Table 5. Multiplace chambers are much larger and are designed to provide treatment to multiple people or to manage complex conditions. Some can even house operating rooms with special precautions. They are typically made of steel with acrylic view ports and are designed for operation up to 6 ATA or 165 feet of sea water. The gauges are reported in feet of sea water on these multiplace chambers to facilitate the use of dive tables for staff or patients. These multiplace chambers are, therefore, best-suited to treat deep water decompression illness. These chambers can accommodate from 2 to 20 people and have variable configurations including horizontal cylinders, spherical shapes, and rectangular chambers.

The primary professional hyperbaric medicine societies in the United States are the Undersea and Hyperbaric Medical Society (UHMS) and the American College of Hyperbaric Medicine. The UHMS has developed a clinical hyperbaric medicine facility accreditation program. This program can be accessed via the UHMS website at http://www.uhms.org, and it was designed to assure that clinical facilities are:

1. Staffed with well-trained specialists;
2. Using high quality equipment that is properly installed, maintained, and operated to the highest possible safety standards;
3. Providing high quality care;
4. Maintaining proper documentation of informed consent, treatment protocols, physician participation, training, and so on (46).

Safety Elements for Equipment and Facilities

The American Society of Mechanical Engineers (ASME) and the Pressure Vessel for Human Occupancy Committee (PVHO) define the design and fabrication guidelines for hyperbaric chambers. Although not required in all states or worldwide, it is accepted as the international standard (46). Next, the National Fire Protection Association (NFPA) has established a safety standard for hyperbaric facilities. The publication, NFPA 99, Safety Standard for Health Care Facilities, Hyperbaric Facilities, Chapter 20 explains the details of and criteria for equipment associated with a hyperbaric chamber facility. The requirements include fire abatement systems, air quality, and electrical requirements. These requirements apply to any hyperbaric chamber placed within a health-care facility. Each site must have a safety director. It is important to have only cotton clothing and to avoid any sources of sparks or static electricity given the 100% oxygen (Fig. 2). In addition to these guidelines, hyperbaric chambers are pressure vessels and, as such, are subject to boiler and pressure vessel laws. They are also medical devices and, in the United States, are also subject to FDA rules for class II medical devices. A chamber is required to have a clearance from the FDA before the device can be legally marketed or distributed, which is often called a 510 k clearance, denoting the form on which the clearance must be submitted. To check on whether a device has received clearance in the United States, one must contact the manufacturer or the Food and Drug Administration (FDA) most easily via their website, http://www.fda.gov/scripts/cdrh/cfdocds/cfpmn/dsearch.cfm.

Facilities must develop defined safety protocols and emergency plans that are available through both the Undersea and Hyperbaric Medicine Society (UHMS) and the American College of Hyperbaric Medicine (ACHM).

FRONTIERS AND INVESTIGATIONAL USES

The use of hyperbaric oxygen therapy has, at times, been surrounded with controversy and spurious claims from improving athletic performance to slowing the aging process. It is essential that the hyperbaric medicine physician, staff, and potential patients understand and follow the principles of evidence-based practice, which means prescribing HBO therapy for the conditions proven to benefit from such treatment. The UHMS website, at www.UHMS. org, and AHCM are good resources for additional information as are numerous publications on hyperbaric medicine such as the hyperbaric medicine textbook available through the UHMS website. Investigational uses for hyperbaric oxygen therapy include carbon tetrachloride poisoning, hydrogen sulfide poisioning, sickle cell crisis, spinal cord injury, closed head injury, cerebral palsy, purpura fulminans, intraabdominal and intracranial abscess, mesenteric thrombosis, retinal artery occlusion, cystoid macular edema, bell's palsy, leprosy, lyme disease, stroke and traumatic brain injury, and brown recluse spider bite. Some of the many investigational uses for HBO therapy may have merit, but these must be rigorously studied using well-designed trials. As the field of hyperbaric medicine continues to advance, so will our understanding of the complex physiologic effects of delivering oxygen under pressure.

AKNOWLEDGMENT

Special thanks to Bob Bartlet, MD, for his help in preparing this chapter.

BIBLIOGRAPHY

1. Bakker DJ, Cramer FS. Hyperbaric surgery. Perioperative care. p 2.
2. Behnke AR, Shaw LA. Use of hyperbaric oxygen in treatment of compressed air illness. Nav Med Bull 1937;35:1–12.
3. Boerma I. High tension oxygen therapy. Proc Royal Soc Med 1964;57(9):817–818.
4. Bakker DJ, Cramer FS. Hyperbaric Surgery Perioperative Care. p. 67.
5. Jain KK. Physical Physiological, and Biochemical Aspects of Hyperbaric Oxygenation. Textbook of Hyperbaric Medicine. Toronto: Hogrefe & Huber Publishers; 1990. p 11.
6. Matos L, Nunez A. Enhancement of healing in selected problem wounds. In: Kindwall EP, Whelan HT, eds. Hyperbaric Medicine Practice, 2nd ed. Flagstaff AZ; Best; 1999.
7. Hunt TK, Pai MP. The effect of varying ambient oxygen tensions on wound metabolism and collagen synthesis. Surg Gynecol Obstet 1972;135:561–567.
8. Knighton DR, Halliday BJ, Hunt TK. Oxygen as an antibiotic: A comparison of the effects of inspired oxygen concentration and antibiotic administration on in vivo bacterial clearance. Arch Surg 1986;121:191–195.
9. Zamboni WA, Roth AC, Russell RC, Graham B, Suchy H, Kucan JO. Morphologic analysis of the microcirculation during reperfusion of ischemic skeletal muscle and the effect of hyperbaric oxygen. Plast Reconstr Surg 1993;91(6):1110–23.
10. Zamboni WA, et al. The effect of acute hyperbaric oxygen therapy on axial pattern skin flap survival when administered during and after total ischemia. J Reconstr Microsurg 1989;5:343–347.
11. Muth CM. Gas embolism (Review). New Eng J Med 342: 476–482.
12. Divers Alert Network: Report on 1994 Diving Accidents. Durham, NC: Duke University; 1995.
13. Barnard EEP, Hanson JM, et al. Post decompression shock due to extravasation of plasma. BMJ 1966;2:154.
14. Powel MR, Spencer MP, Von Ramm OT. Ultrasonic surveillance of decompression. In: Bennett PB, Elliott DH, eds. The Physiology of Diving and Compressed Air Work, 3rd ed. London Bailliere: Tindall; 1982. pp 404–434.
15. Zamboni WA. Applications of hyperbaric oxygen therapy in plastic surgery. In: Oriani G, Marroni A, eds. Handbook on Hyperbaric Medicine, 1st ed. New York: Springer; 1995. p. 443–484.
16. Hunt TK. The physiology of wound healing. Ann Emerg Med 1988;17:1265–1273.
17. Hunt TK, Hopf HW. Wound healing and wound infection. Surg Clinics N Am 1997;77(3):587–606.
18. Oriani G, Micheal M, Meazza D, et al. Diabetic foot and hyperbaric oxygen therapy: A ten-year experience. J Hyperbar Med 1992;7:213–221.
19. Wattel FE, Mathieu DM, Fossati P, et al. Hyperbaric oxygen in the treatment of diabetic foot lesions: Search for healing predictive factors. J Hyperbar Med 1991;6:263–267.
20. Baroni G, Porro T, Faglia E, Pizzi G, et al. Hyperbaric oxygen in diabetic gangrene treatment. Diabetes Care 1987;10:81–86.
21. Oriani G, Micheal M, Meazza D, et al. Diabetic foot and hyperbaric oxygen therapy: A ten-year experience. J Hyperbar Med 1992;7:213–221.
22. Strauss MB, Bryant BJ, Hart GB. Transcutaneous oxygen measurements under hyperbaric oxygen conditions as a predictor for healing of problem wounds. Foot Ankle Int. 2002 Oct; 23(10):933–7.
23. Zamboni WA, Roth AC, Russel RC, Nemiroff PM, Casas L, Smoot EC. Hyperbaric oxygen improves axial skin flap survival when administered during and after total ischemia. J Reconstr Micro 1989;5:343–347.
24. Hunt TK, Pai MP. The effect of varying ambient oxygen tensions on wound metabolism and collagen synthesis. Surg Gyn Obstet 1972;135:561–567.
25. Niinikoski J, Hunt TK. Oxygen Tension in Human Wounds. J Surg Res 1972;12:77–82.
26. Mathieu D, et al. Pedicle musculocutaneous flap transplantation: prediction of final outcome by transcutaneous oxygen measurements in hyperbaric oxygen. Plast Reconstr Surg 1993;91:329–334.
27. Waterhouse MA, et al. The use of HBO in compromised free tissue transfer and replantation: A clinical review. Undersea Hyperb Med 1993;20(Suppl):54 (Abstract).
28. Perrins DJD, Cantab MB. Influence of hyperbaric oxygen on the survival of split skin grafts. Lancet 1967;1:868–871.
29. Persons BL, Zamboni WA. Hyperbaric oxygen in plastic and reconstructive surgery. In: Bakker DJ, Cramer FS, eds. Hyperbaric Surgery Perioperative Care. Flagstaff, AZ: Best; 2002.
30. Mazelowski MC, Zamboni WA, Haws MF, Smoot EC, Stephenson LL. Effect of hyperbaric oxygen on composite graft survival in a rat ear model. Undersea and Hyperbaric Med Suppl 1995;22:50.
31. McFarlane RM, Wermuth RE. The use of hyperbaric oxygen to prevent necrosis in experimental pedicle flaps and composite skin grafts. Plast Reconstr Surg 1966; 37:422–430.
32. Angel MF, et al. Free radicals: Basic concepts concerning their chemistry, pathophysiology, and relevance to plastic surgery. Plast Reconstr Surg 79:990.
33. Lozano DD, Zamboni WA, Stephenson LL. Effect of hyperbaric oxygen and medicinal leeching on survival of axial skin flaps subjected to total venous occlusion. Undersea Hyperb Med suppl 1997;24:86.
34. Thom SR, Bhopale V, Fisher D, Manevich Y, Huang PL, Buerk DG. Stimulation of nitric oxide synthase in cerebral cortex due to elevated partial pressures of oxygen: an oxidative stress response. J Neurobiol 2002;51(2):85–100.

35. Jones S, Wang WZ, Natajaraj C, Khiabani, Stephenson LL, Zamboni WA. HBO inhibits IR induced Neutrophil CD 18 Polarization by a nitric oxide mechanism. Undersea Hyperb Med 2002;35 (Suppl):75.
36. Zamboni WA, Roth AC, Russel RC, Nemiroff PM, Casas L, Smoot EC. Hyperbaric oxygen improves axial skin flap survival when administered during and after total ischemia. J Reconstr Micro 1989;5:343–347.
37. Gimbel M, Hunt TK. Wound healing and hyperbaric oxygen. In: Kindwall EP, Whelan HT, eds. Hyperbaric Medicine Practice, 2nd ed. Flagstaff, AZ: Best; 1999. p 169–204.
38. Bartlett B. Hyperbaric therapy. Radiation Injury 1994:2–3. HBO has been shown to increase angiogenesis and blood flow in previously irradiated tissue or bone. (Marx RE, Ehler WJ, Taypongsak PT, Pierce LW. Relationship of oxygen dose to angiogenesis induction in irradiated tissue. Am J Surg 1990; 160:519–524.
39. Carl UM, Feldmeier JJ, Schmitt G, Hartmann KA. Hyperbaric oxygen therapy for late sequelae in women receiving radiation after breast conserving surgery. Int J Radiat Oncol Biol Phys 2001;49:1029–1031.
40. Feldmeier JJ. Hyperbaric oxygen: Does it have a cancer causing or growth enhancing effect. Proceedings of the consensus conference sponsored by the European society for therapeutic radiology and oncology and the European committee for hyperbaric medicine. Portugal, 2001:129–146.
41. Gray KH, Conger AD, Ebert M, Hornsey S, Scott OCA. The concentration of oxygen dissolved in tissues at the time of irradiation as a factor in radiotherapy. Br J Radiol 1953;26: 638–648.
42. Wada J, Ikeda T, Kamata K, Ebuoka M. Oxygen hyperbaric treatment for carbon monoxide poisoning and severe burn in coal mine gas explosion. Igakunoaymi (Japan) 1965;54–68.
43. Germonpre P, Reper P, Vanderkelen A. Hyperbaric oxygen therapy and piracetam decrease the early extension of deep partial thickness burns. Burns 1996;6:468–473.
44. Cianci P, Sato R, Green B. Adjunctive hyperbaric oxygen reduces length of hospital stay, surgery and the cost of care in severe burns. Undersea Biomed Research Suppl 1991;18:108.
45. Bartlett R. Carbon monoxide poisoning. In: Haddad M, Shannon M, Winchester J, eds. Poisoning and Drug Overdose, 3rd ed. New York: WB Saunders Company; 2002.
46. Bakker DJ, Cramer FS. Hyperbaric Surgery Perioperative Care. Flagstaff, AZ: Best Publishing; 2002.

See also BLOOD GAS MEASUREMENTS; HYPERBARIC OXYGENATION; OXYGEN MONITORING; PULMONARY PHYSIOLOGY; RESPIRATORY MECHANICS AND GAS EXCHANGE; VENTILATORY MONITORING.

HYPERBARIC OXYGENATION

HARRY T. WHELAN
JEFFREY A. NIEZGODA
MATTHEW C. LEWIS
Medical College of Wisconsin
Milwaukee, Wisconsin

ERIC P. KINDWALL
BERNADETTE CABIGAS
St. Luke's Medical Center
Milwaukee, Wisconsin

INTRODUCTION

Hyperbaric oxygen (HBO) is simply the delivery of molecular oxygen in very high dosage. Even though experience has shown HBO to be very useful in a number of conditions, the exact mechanism of action at the molecular level is not fully understood. Studies done by Thom et al. (1) demonstrated that elevated oxygen tensions stimulated neuronal nitric oxide synthase (NOS1) and increased steady-state nitric oxide concentration in their microelectrode-implanted rodents. Buras et al. (2) in their studies with human umbilical vein endothelial cells (HUVEC) and bovine aortic endothelial cells (BAEC) showed that hyperbaric oxygen (HBO) down-regulated intracellular adhesion molecule 1 (ICAM 1) expression via the induction of endothelial nitric oxide synthase (NOS3), which proved beneficial in treating ischemia reperfusion injuries. Other studies talk about interactions between nitric oxide and oxygen species and their role in various disease states. Clearly, interest in HBO is growing.

Boerema (3) introduced hospital use of the hyperbaric chamber in the late 1950s in Holland, simply to maintain a semblance of normoxia in patients undergoing cardiac surgery. Heart–lung machines had not yet been invented, and the use of the chamber made certain kinds of cardiac surgery possible for the first time. Boerema felt that if enough oxygen could be driven physically into solution in the tissues, which he termed "drenching", the circulation to the brain could be interrupted longer than 3–4 min. It also rendered surgery on many pediatric patients less risky. For example, if the normal arterial pO_2 in a patient with Tetralogy of Fallot was 38 mmHg, placing him in the chamber might raise it to 94 mmHg. Operating on the patient under hyperbaric conditions posed much less risk of ventricular fibrillation when the heart or great vessels were manipulated.

This idea caught on quickly, and soon large surgical hyperbaric chambers were built in Glasgow, New York, Los Angeles, Chicago, Minneapolis, and at Boston Children's Hospital. By the early 1960s, however, heart–lung machines became more common, and the need to do surgery in the hyperbaric chamber diminished substantially. Many large surgical chambers were left to gather dust or were dismantled, as hospital floor space is always at a premium. During this time the surgeons, who had been doing most of the research, left the field. Of the nondiving conditions, only carbon monoxide poisoning and gas gangrene seemed to be likely candidates for hyperbaric oxygen treatment based on credible research.

In 1969, however, a double-blind controlled study on the use of hyperbaric oxygen in senility was published in *The New England Journal of Medicine*. Results seemed promising, and this initiated the propagation of hyperbaric quackery. The original investigators made no sweeping claims for the research, but simply felt that the area merited further investigation. Eventually, further research showed that the results of the study reported in the *New England Journal* article were a statistical anomaly and could not be reproduced. However, hyperbaric enthusiasts seized upon the earlier report, and senility began to be treated in hyperbaric chambers, along with a host of other diseases. Most of these were not in medical centers. Fly-by-night "clinics" suddenly appeared claiming to cure anything and everything. Patients were treated for skin wrinkles, loss of sexual vigor, and a host of other maladies. As there were few investigators doing good research in the area at that time, the field fell into disrepute.

Fortunately, a few legitimate investigators persisted in their work, looking at the effects of hyperbaric oxygen in greater detail. Soon it became clear that under hyperbaric conditions oxygen had some unusual effects. The Undersea and Hyperbaric Medical Society created a committee to investigate the field. After careful study, the committee laid down guidelines for what should be reimbursed by third-party payers and what conditions should be considered investigational. Their report appeared in 1977 and was adopted as a source document for Blue Cross/Blue Shield (4). About the same time, Jefferson C. Davis of the United States Air Force School of Aerospace Medicine edited the first textbook in hyperbaric medicine (5). It was only then that a firm scientific basis was reestablished for the field, leading to increased acceptance by the medical community. The number of chambers operating in hospitals has risen dramatically from only 37 in 1977 to > 500 today. The Undersea and Hyperbaric Medical Society (www.UHMS.org) and the American College of Hyperbaric Medicine (www.ACHM.org) have taken responsibility for setting standards in this field and for encouraging additional research. At this time, ~ 13 clinical disorders have been approved for hyperbaric treatment. They include air or gas embolism, carbon monoxide poisoning, clostridial myonecrosis, crush injury or compartment syndrome, decompression sickness, problem wounds, severe blood loss anemia, necrotizing soft tissue infections, osteomyelitis, radiation tissue damage, skin grafts or flaps, thermal burns and brain abscess.

Remember that hyperbaric oxygen was introduced initially into hospitals in order to simply maintain normoxia or near-normoxia in patients undergoing surgery. It was only later, and quite serendipitously that researchers discovered that oxygen under increased atmospheric pressure gained some of the attributes of a pharmacologic agent. Oxygen begins to act like a drug when given at pressures of 2 atm or greater. For example, oxygen under pressure can terminate lipid peroxidation *in vivo* (6), it can enhance the bacteriocidal capabilities of the normal leukocyte (7,8), and it can stimulate the growth of new capillaries in chronically ischemic tissue, such as in the diabetic foot, or in tissue that has undergone heavy radiation. It can reduce intracranial pressure on the order of 50% within seconds of its initiation, and this effect is additive to that of hypocapnia (9–11). HBOT can increase the flexibility of red cells, augmenting the effects of pentoxifylline (12). It can decrease edema formation by a factor of 50% in postischemic muscle and prevent second-degree burn from advancing to full-thickness injury (13–15). Hyperbaric oxygen has also been shown to hasten functional recovery of traumatized peripheral nerves by almost 30% following repair. Many of these discoveries have been made only in the last decade.

In a number of these areas, we are beginning to understand the basic mechanisms of action, but overall very little is understood at the molecular level. It is anticipated that studies involving nitric oxide synthase will provide insight regarding the elusive molecular mechanistic explanation. Also, many contributions to our understanding have come from advances made in the biochemistry of normal wound healing. We understand that normal oxygen pressures are 80–90-mmHg arterially, that oxygen enters our tissues from the capillaries, and that at this interface carbon dioxide (CO_2) is removed. Under hyperbaric conditions, all of this changes. At a chamber pressure of 2.4 atm (ATA), the arterial oxygen pressure (pO_2) reaches ~ 1500 mmHg, immediately saturating the red blood cells (RBCs). Upon reaching the tissues, these RBCs never unload their oxygen. At this high partial pressure of gas, oxygen diffuses into the tissues directly from the plasma. Returning to the heart, the RBCs are bathed in plasma with a pO_2 of 150–200 mmHg. Tissue oxygen requirements are completely derived from the plasma. In theory, one might think that this condition could prove fatal, as red cells no longer can carry CO_2 away from the tissues. However, we are fortunate that CO_2 is 50 times more soluble in plasma than are oxygen and nitrogen, and the body has a very capable buffering system which overcomes the loss of the Haldane effect, which is the increase in CO_2 carrying capacity of deoxygenated red cells (16).

Another factor to be considered is the actual part of the circulatory system that overcomes the loss of the Haldane effect. Traditionally, we think of this exchange occurring in the capillaries. Under very high pressures, however, computer modeling has shown that nitrogen exchange under pressure (as in deep sea divers) is probably complete by the time the blood reaches the arteriolar level. Whether this is true when hyperbaric oxygen is breathed has not yet been determined. The rate of metabolism under hyperbaric conditions appears to be unchanged, and the amount of CO_2 produced appears to be about the same as when breathing air. It would be interesting to know just at what level oxygen exchange is accomplished in the tissues, as this might have practical implications when treating people with severe capillary disease.

Oxygen can be toxic under pressure. Pulmonary toxicity and lung damage can be seen at oxygen pressures > 0.6 atm during chronic exposure. Central nervous system (CNS) toxicity can manifest as generalized seizure activity when oxygen is breathed at pressures of 3 atm or greater. The CNS toxicity was first observed by Paul Bert in 1878, and is termed the "Paul Bert Effect" (17). Despite years of research into this phenomenon, the exact underlying or molecular cause of the seizure has not yet been discovered. There is a generalized vasoconstriction that occurs when oxygen is breathed at high pressure, reducing blood flow to muscle, heart, and brain by a factor of $\sim 20\%$, as a defense against toxic quantities of oxygen. The exact mechanism responsible for this phenomenon is not fully understood.

Central nervous system oxygen toxicity was evaluated by the Royal Navy. The purpose of this research was to determine the time until convulsion so that combat swimmers would know their endurance limits under various conditions. Volunteer research subjects swam in a test tank using closed-circuit oxygen rigs until convulsion occurred and thus established safe oxygen tolerance boundaries.

Also related to the effect of oxygen, the "off" phenomenon (18) was first described by Donald in 1942. He observed that seizures sometimes occurred when the chamber pressure was reduced or when a diver surfaced and oxygen breathing under pressure was suddenly terminated. Lambertsen (19) provided a description of this type of seizure activity:

> The convulsion is usually but not always preceded by the occurrence of localized muscular twitching, especially about the eyes, mouth and forehead. Small muscles of the hands may also be involved, and incoordination of diaphragm activity in respiration may occur. After they begin, these phenomena increase in severity over a period which may vary from a few minutes to nearly an hour, with essentially clear consciousness being retained. Eventually an abrupt spread of excitation occurs and the rigid tonic phase of the convulsion begins. Respiration ceases at this point and does not begin again until the intermittent muscular contractions return. The tonic phase lasts for about 30 seconds and is accompanied by an abrupt loss of consciousness. It is followed by vigorous clonic contractions of the muscle groups of the head and neck, trunk and limbs. As the incoordinated motor activity stops, respiration can proceed normally.

Within the wound healing community, current doctrine holds that a tissue pO_2 of 30–40 mmHg is necessary for adequate wound healing (20,21). Below 30 mmHg, fibroblasts are unable to replicate or produce collagen. Additionally, when the pO_2 drops < 30 mmHg, leukocytes are unable to utilize oxidative mechanisms to kill bacteria. We have noted that the tissue pO_2 is critical, but that the actual quantity of oxygen consumed in wound healing is relatively small. The amount of oxygen used to heal a wound is only $\sim 10\%$ of that required for brain metabolism.

Production of new collagen is also a requirement for capillary ingrowth or proliferation (22). As capillaries advance, stimulated by angiogenic growth factor, they must be supported by an extracellular collagen matrix to facilitate ingrowth into tissue. In the absence of new collagen, capillary ingrowth cannot occur. This effect is crucial in treating radionecrosis (23–25), where the tissue is primarily hypovascular, and secondarily hypoxic and hypocellular. It has been discovered that when collagen production can be facilitated, new capillaries will invade the previously irradiated area, and healing will then occur. The tissue pO_2 rises to $\sim 80\%$ of normal and plateaus; however, this is sufficient for healing and will even support bone grafting. Historically, the only means of managing radionecrosis was to excise the radiated area and bring in fresh tissue with its own blood supply. New collagen formation and capillary ingrowth also account for the rise in tissue pO_2, which can be achieved in patients with diabetic foot lesions.

It is now well understood that the stimulus for growth factor production by the macrophage is hypoxia and/or the presence of lactic acid (26,27). Wounds managed in hyperbaric units are typically ischemic and hypoxic. Periods of relative hypoxia, required for the stimulation of growth factor production, exist between hyperbaric treatments.

Surprisingly, oxygen levels remain high in tissues for longer than one would expect following hyperbaric treatment. In a study by George Hart (28) at Long Beach Memorial Hospital, a mass spectrometer probe was inserted in the unanesthetized thigh tissues of normal volunteers. Muscle and subcutaneous tissue pO_2 values in study subjects remained significantly elevated for 2–3 h following hyperbaric oxygen treatment. Arterial pO_2 was also measured and found to rise immediately and significantly under hyperbaric conditions but returned to normal levels within a couple of minutes upon egress from the chamber (Fig. 1). Thus, multiple daily HBO treatments can maintain useful oxygen levels for up to 12 h/day.

Mention has been made of enhanced leukocyte killing of bacteria under hyperbaric conditions. Jon Mader of the University of Texas-Galveston (29) carried out a rather simple, but elegant, experiment to demonstrate this. The fascinating part of this study is that in the evolution of the human body, a leukocyte has never been exposed to a partial pressure of 150 mmHg while in tissues. This level is impossible to attain breathing air. Nevertheless, when one artificially raises the pO_2 far beyond the leukocyte's normal functional parameters, it becomes even more lethal to bacteria. This is an anomaly, as one rarely can improve on Mother Nature. Of some interest in this regard is that if one bites one's tongue, one is never concerned about possible infection, even though it is a human bite. Similarly, hemorrhoidectomies rarely, if ever, become infected. The reason is that the pO_2 of the tissues in and around the oral cavity are very high, and the pO_2 in hemorrhoidal veins is nearly arterial. Tom Hunt has shown it is impossible to infect tissue that is injected with raw staphylococci if the pO_2 in the same tissue is > 50 mmHg. Both he and David Knighton have described oxygen as an antibiotic (30,31).

The reduction of intracranial pressure is facilitated by vasoconstriction. Experimentally, Rockswold has shown that mortality can be halved in victims of closed head injury with Glasgow Coma Scales in the range of 4–6. One of the major mechanisms here is a reduction of intracranial pressure while continuing to oxygenate hypoxic brain (32–36). Sukoff et al. (37) administered 100% O_2, 1.5 ATA × 60 min every 24 h (maximum of 7 h) to severely brain injured patients. This resulted in a 50% reduction in mortality.

A paper published by Mathieu (38) has shown that the flexibility index of red cells can be changed from 23.2 to 11.3 within 15 hyperbaric treatments. This increase in flexibility can prove quite useful in people with narrowed capillaries. However, whether this phenomenon plateaus at 15 treatments, its duration and underlying mechanism are still unknown.

Nylander et al. (39) demonstrated that following complete occlusion of the blood flow to rat leg for 3 h, postischemic edema could be reduced by 50% if the animals are promptly treated with hyperbaric oxygen. He also demonstrated that the mechanism for this was preservation of adenosine triphosphate (ATP) in the cells, which provides the energy for the cells to maintain their osmolarity. Cianci

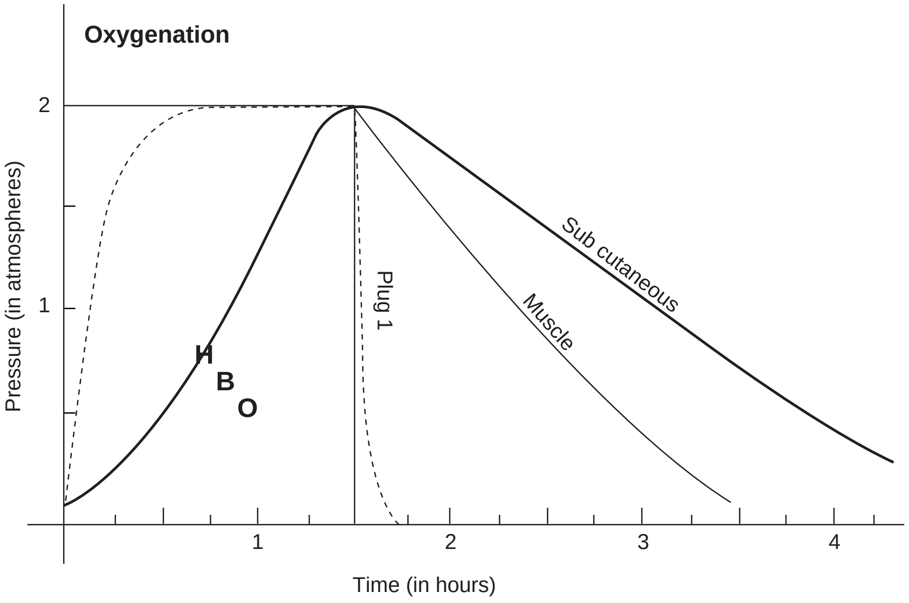


Figure 1. Arterial, muscle and subcutaneous pO_2 after HBO treatment.

(40,41). Yamaguchi, and others have underscored the importance of ATP in preventing edema in burn. Treatment twice daily has shown to be more efficacious than treatment once a day.

Zamboni (42) and Gingrass, working at the University of Southern Illinois, have shown the effects of hyperbaric oxygen on speeding functional return in peripheral nerve repair and grafting. At 6 weeks, there is a 28% improvement of function in the affected leg of these rats.

Niezgoda (43) performed a randomized and double-blinded study in human volunteers investigating the effect of hyperbaric oxygen in a controlled burn wound model. He demonstrated statistically significant decreases in edema formation and wound exudate in the hyperbaric oxygen treated group.

Finally, the mechanism for the effects of hyperbaric oxygen in carbon monoxide poisoning (44,45) is now better understood. Traditionally, it was felt that the mere presence of carboxyhemoglobin blocked transport of oxygen to the tissues. However, studies by Goldbaum et al. (46) at the Armed Forces Institute of Pathology in 1975 lead us to different conclusions. Impairment of cytochrome *A*3 oxidase and lipid peroxidation occurring following a reperfusion injury are now suggested as the primary pathways in the pathophysiology causing fatality. Stephen Thom (47–50) at the University of Pennsylvania pioneered research in this area. It appears that as carbon monoxide levels fall, the products of lipid peroxidation rise, indicating that brain damage is occurring only during the resuscitative phase, thus becoming reperfusion injury. Thom demonstrated that a period of hypotension (even though it may only be a matter of seconds) is enough to initiate lipid peroxidation. Oxygen at 1 atm has no effect on halting the process. However, oxygen at 3 atm terminates lipid peroxidation. Patients who have been treated acutely with hyperbaric oxygen rarely exhibit signs of delayed deterioration, reported in 30–40% of severe cases treated only with normobaric oxygen. The probable mechanism for this is the ability of hyperbaric oxygen at three ATA to terminate lipid peroxidation.

Finally, in many ways it seems paradoxical that oxygen at high pressure, which intuitively would seem to provide more substrate for free-radical formation, still benefits tissues from crush injury and postischemic states. But ironically, it is precisely this hyperbaric environment that promotes certain so-called reactive oxygen species with inherent protective qualities (51). More studies are certainly needed to investigate the underlying pharmacologic benefits afforded by hyperbaric oxygen. We have only just begun to explore and utilize a treatment modality whose time has come.

BIBLIOGRAPHY

1. Thom SR, Fisher D, Zhang J, Bhopale VM, Ohnishi ST, Kotake Y, Ohnishi T, Buerk DG. Stimulation of perivascular nitric oxide synthesis by oxygen. Am J Physiol Heart Circ Physiol 2003;284:H1230–1239.
2. Buras JA, Stahl GL, Svoboda KKH, Reenstra WR. Hyperbaric oxygen downregulates ICAM-1 expression induced by hypoxia and hypoglycemia: the role of NOS. Am J Physiol Cell Physiol 2000;278:C292–C302.
3. Boerema I, Kroll JA, Meijne NG, Lokin E, Kroon B, Huiskes JW. High atmospheric pressure as an aid to cardiac surgery. Arch Chir Neerl 1956;8:193–211.
4. Kindwall EP. Report of the committee on hyperbaric oxygenation. Bethesda: Undersea Medical Society; Bethesda 1977.
5. Davis JC, Hunt TK. Hyperbaric oxygen therapy. Undersea Medical Society; Bethesda 1977.
6. Thom SR. Molecular mechanism for the antagonism of lipid peroxidation by hyperbaric oxygen. Undersea Biom Res (Suppl) 1990;17:53–54.
7. Andersen V, Hellung-Larsen P, Sorensen SF. Optimal oxygen tension for human lymphocytes in culture. J Cell Physiol 1968;72:149–152.
8. Gadd MA, McClellan DS, Neuman TS, Hansbrough JF. Effect of hyperbaric oxygen on murine neutrophil and T-lymphocyte functions. Crit Care Med 1990;18:974–979.

9. Hayakawa T, Kanai N, Kuroda R, et al. Response of cerebrospinal fluid to hyperbaric oxygenation. J Neurol Neurosurg Psych 1971;34:580–356.
10. Miller JD, Fitch W, Ledingham IM, et al. Reduction of increased intracranial pressure. Neurosurgery 1970;33: 287–296.
11. Sukoff MH, Ragatz RE. Hyperbaric oxygenation for the treatment of acute cerebral edema. Neurosurgery 1982;10: 29–38.
12. Nemiroff PM. Synergistic effects of pentoxifylline and hyperbaric oxygen on skin flaps. Arch Otolaryngol Head Neck Surg 1988;114:977–981.
13. Cianci P, Lueders H, Shapiro R, Sexton J, Green B. Current status of adjunctive hyperbaric oxygen in the treatment of thermal wounds. In: Proceedings of the second Swiss symposium on hyperbaric medicine; Baker DJ, JS, editors. Foundation for Hyperbaric Medicine; Basel: 1988. p 163–172.
14. Grossman AR. Hyperbaric oxygen in the treatment of burns. Ann Plast Surg 1978;1:163–171.
15. Hart GB, O'Reilly RR, Broussard ND, Cave RH, Goodman DB, Yanda RL. Treatment of burns with hyperbaric oxygen. Surg Gynacol Obstet 1974;139:693–696.
16. Coburn RF, Forster RE, Kane PB. Considerations of the physiological variables that determine the blood carboxyhemoglobin concentrations in man. J Clin Invest 1965;44: 1899–1910.
17. Bert P. Barometric Pressure. 1879. p 579. (translated by Hitchcock MS, Hitchcock FA) Bethesda: Reprinted by the Undersea and Hyperbaric Medicine Society; 1978.
18. Donald KW. Oxygen poisoning in man. Br Med J 1947; 712–717.
19. Lambertsen CJ. In: Fenn WO, Rahn H, editors. Handbook of Physiology, Respiration. Washington, D.C.: American Physiological Society; Section 3, Volume II. p 1027–1046.
20. Knighton DR, Hunt TK, Scheuenstuhl H, Halliday B, Werb Z, Banda MJ. Oxygen tension regulates the expression of angiogenesis factor by macrophages. Science 1983;221:1283–1285.
21. Knighton DR, Oredsson S, Banda MJ, Hunt TK. Regulation of repair: hypoxic control of macrophage mediated angiogenesis. In: Hunt TK, Heppenstall RB, Pines E, Rovee D, editors. Soft and hard tissue repair. New York: Praeser; 1948. p 41–49.
22. Knighton DR, Hunt TK, Thakral KK, Goodson WH. Role of platelets and fibrin in the healing sequence, an in vivo study of angiogenesis and collagen synthesis. Ann Surg 1982;196: 379–388.
23. Davis JC. Soft tissue radiation necrosis: The role of hyperbaric oxygen. HBO Rev 1987;2(3):153–167.
24. Davis JC, et al. Hyperbaric oxygen: A new adjunct in the management of radiation necrosis. Arch Otol 1979;105:58–61.
25. Hart GB, Mainous EG. The treatment of radiation necrosis with hyperbaric oxygen (OHP). 1976;37:2580–2585.
26. Jensen JA, Hunt TK, Scheuenstuhl H, Banda MJ. Effect of lactate, pyruvate, and physician on secretion of angiogenesis and mitogenesis factors by macrophages. Lab Invest 1986;54: 574–578.
27. Knighton DR, Schumerth S, Fiegel VD. Microenvironmental regulation of macrophage growth factor production. In preparation.
28. Hart GB, Wells CH, Strauss MB. Human skeletal muscle and subcutaneous tissue carbon dioxide, nitrogen and oxygen gas tension measurement under ambient and hyperbaric conditions. J App Res Clin Exper Therap Spring 2003;3(2).
29. Wang J, Corson K, Mader J. Hyperbaric oxygen as adjunctive therapy in vibrio vulnificus septicemia and cellulites. Undersea Hyperbaric Med Spring 2004, 31(1):179–181.
30. Hunt TK, Linsey M, Grislis G, Sonne M, Jawetz E. The effect of differing ambient oxygen tensions on wound infections. Ann Surg 1975;181:35–39.
31. Knighton DR, Halliday B, Hunt TK. Oxygen as an antibiotic. A comparison of the effects of inspired oxygen concentration and antibiotic administration on in vivo bacterial clearance. Arch Surg 1986;121:191–195.
32. Miller JD, et al.The effect of hyperbaric oxygen on experimentally increased intracranial pressure. J Neurosurg 1970;32: 51–54.
33. Miller JD, Ledingham IM. Reduction of increased intracranial pressure: Comparison between hyperbaric oxygen and hyperventilation. Arch Neurol 1971;24:210–216.
34. Mogami H, et al.Clinical application of hyperbaric oxygenation in the treatment of acute cerebral damage. J Neurosurg 1969;31:636–643.
35. Sukoff MH, et al.The protective effect of hyperbaric oxygenation in experimental cerebral edema. J Neurosurg 1968;29: 236–241.
36. Sukoff MH, Ragatz RE. Hyperbaric oxygen for the treatment of acute cerebral edema. Neurosurgery 1982;10(1):29–38.
37. Sukoff MH. Effects of hyperbaric oxygenation [comment]. J Neurosurg 2001;94(3):403–411.
38. Mathieu D, Coget J, Vinkier L, Saulnier F, Durocher A, Wattel F. Red blood cell deformability and hyperbaric oxygen therapy. (Abstract) HBO Rev 1985;6:280.
39. Nylander G, Lewis D, Nordstrom H, Larsson J. Reduction of postischemic edema with hyperbaric oxygen. Plast Reconstr Surg 1985;76:596–601.
40. Cianci P, Lueders HW, Lee H, Shapiro RL, Sexton J, Williams C, Green B. Adjunctive hyperbaric oxygen reduces the need for surgery in 40-80% burns. J Hyper Med 1988;3: 97–101.
41. Cianci P, Lueders HW, Lee H, Shapiro RL, Sexton J, Williams C, Green B. Hyperbaric oxygen and burn fluid requirements: Observations in 16 patients with 40-80% TBSA burns. Undersea Biomed Res (Suppl) 1988;15:14.
42. Zamboni WA, Roth AC, Russell RC, Nemiroff PM, Casa L, Smoot C. The effect of acute hyperbaric oxygen therapy on axial pattern skin flap survival when administered during and after total ischemia. J Reconst Microsurg 1989;5: 343–537.
43. Niezgoda JA, Cianci P. The effect of hyperbaric oxygen on a burn wound model in human volunteers. J Plast Reconstruct Surg 1997;99:1620–1625.
44. Brown SD, Piantadosi CA. Reversal of carbon monoxide-cytochrome C oxidase binding by hyperbaric oxygen in vivo. Adv Exp Biol Med 1989;248:747–754.
45. End E, Long CW. Oxygen under pressure in carbon monoxide poisoning. J Ind Hyg Toxicol 1942;24:302–306.
46. Goldblum LR, Ramirez RG, Absalon KB. Joint Committee on Aviation Pathology XII. What is the mechanism of carbon monoxide toxicity? Aviat Space Environ Med 1975;46(10): 1289–1291.
47. Thom SR. Antagonism of carbon monoxide-mediated brain lipid peroxidation by hyperbaric oxygen. Toxicol Appl Pharmacol 1990;105:340–344.
48. Thom SR, Elbuken ME. Oxygen-dependent antagonism of lipid peroxidation. Free Rad Biol Med 1991;10:413–426.
49. Thom SR. Carbon-monoxide mediated brain lipid peroxidation in the rat. J Appl Physiol 1990;68:997–1003.
50. Thom SR. Dehydrogenase conversion to oxidase and lipid peroxidation in brain after carbon monoxide poisoning. J Appl Physiol 1992;73:1584–1589.
51. Thom SR, Bhopale V, Fisher D, Manevich Y, Huang PL, Buerk DG. Stimulation of nitric oxide synthase in cerebral cortex due to elevated partial pressures of oxygen: An oxidative stress response. J Neurobiol 2002;51:85–100.

See also HYBERBARIC MEDICINE; OXYGEN MONITORING.

HYPERTENSION. See BLOOD PRESSURE MEASUREMENT.

HYPERTHERMIA, INTERSTITIAL

MICHAEL D. SHERAR
London Health Sciences Centre
and University of Western
Ontario
London, Ontario, Canada
LEE CHIN
University of Toronto
Toronto, Ontario, Canada
J. CARL KUMARADAS
Ryerson University
Toronto, Ontario, Canada

INTRODUCTION

Interstitial hyperthermia or thermal therapy is a minimally invasive method for the treatment of cancer. Radio frequency (RF), microwave, laser light, or ultrasound energy is delivered through one or more thin needle devices inserted directly into the tumor.

Interstitial devices have the significant advantage over external devices of being able to deliver thermal energy directly into the target region, thereby avoiding depositing energy into intervening nontarget tissue. Their main disadvantage is that the needle devices employed often deposit energy over only a small volume. This can make it challenging to deliver an adequate thermal dose to large target regions. This problem was highlighted in an early radiation therapy oncology group (RTOG) phase III trial in which only 1 out of 86 patients was deemed to have received an adequate thermal treatment (1).

These early challenges in interstitial hyperthermia have been addressed, to some extent, through the development of improved heating devices and more detailed monitoring of applicator placement and dose delivery. Quality assurance guidelines have been developed by the RTOG to raise the quality of heating (2). The guidelines recommend pretreatment planning and equipment checks, the implantation of considerations and documentation, the use of thermometry, and the development of safety procedures. Treatment procedures have also been improved through the use of more detailed thermometry, especially using magnetic resonance imaging approaches (3,4).

THERMAL DOSE AND HEAT TRANSFER

The goal of interstitial thermal therapy is to deliver a prescribed dose to a target volume. Thermal dose is defined as equivalent minutes at 43 °C, or TD. The units of TD are minutes, which represents the time tissue would need to be maintained at a constant temperature of 43 °C to have the same effect as the particular time–temperature history that the tissue was exposed to. The thermal dose after □ minutes of heating can be calculated if the time–temperature history is known (5),

$$\mathrm{TD}(t) = \int_0^t R^{43-T(\tau)} d\tau \qquad \text{where}$$

$$R = \begin{cases} 0.25 \text{ for } T \le 43\,^\circ\mathrm{C} & (1) \\ 0.5 \text{ for } T > 43\,^\circ\mathrm{C} & (2) \end{cases}$$

The dose prescribed for treatment depends on whether the heating is being used as an adjuvant to radiation or systemic therapy, or whether it is being used as a stand-alone treatment to coagulate tissue. For the former use, the dose prescribed is typically 10–60 min (Eq. 1) and for the latter it is usually prescribed to be > 240 min (Eq. 2). This is because temperatures employed for adjuvant treatment (usually referred to as hyperthermia) are in the 40–45 °C range. For stand-alone coagulation (usually referred to as thermal therapy or thermal ablation), temperatures in the range of 55–90 °C are used.

The temperature (T) produced in tissue depends on the heat deposition by the applicator, heat conduction, and blood flow according to

$$\rho c \frac{\partial T}{\partial t} - \nabla \cdot (k \nabla T) + \mathbf{v} \cdot \nabla T = Q$$

where ρ is the tissue mass density, ∇ is the heat capacity of the tissue, k is the thermal conductivity of the tissue, $\mathbf{v}$ is the blood velocity profile, and Q is the heat absorbed per unit volume. Detailed knowledge of the blood velocity profile at the capillary level is generally unknown, and even if it were known the calculations would require impractically large computational resources. While several models have been proposed to calculate heat transfer due to perfusion, the Pennes bioheat transfer equation is most often employed (6)

$$\rho c \frac{\partial T}{\partial t} - \nabla \cdot (k \nabla T) + w c_{\mathrm{b}} (T - T_{\mathrm{b}}) = Q$$

where w is blood mass perfusion rate, c_{b} is the blood heat capacity, and T_{b} is the temperature of the blood entering the treatment field, and $\mathbf{v}$ is the velocity field of any convective flow (e.g., as the blood in large vessels). This equation can be used to predict the temperature in tissue, and therefore plan thermal therapy or hyperthermia treatments if the perfusion rate is known. Penne's equation does not accurately predict for the effect of large blood vessels that must be modeled individually.

ELECTROMAGNETIC HEATING

The heat absorbed (or deposited) in tissue is often described in terms of the power per unit mass. It is called the specific absorption rate or SAR. For electromagnetic devices heat is deposited by the motion of charges or ions. The movement of charge depends on the electric field produced by the applicator in tissue. In microwave and RF hyperthermia, the applicators are driven by sinusoidally time-varying signals. In this case, the electric field can be written in phasor form $\mathbf{E}$ such that the electric field is given by, $E(t) = \Re(\mathbf{E}e^{j\omega t})$, where $\Re(\mathbf{x})$ is the real part of the complex vector $\mathbf{x}$, and ω is the angular frequency of the driving signal. The SAR is then

$$\mathrm{SAR} = \frac{Q}{\rho} = \frac{\sigma}{2\rho}(\mathbf{E} \cdot \mathbf{E}^{*})$$

where σ is the electrical conductivity of the tissue.

The calculation of the electric field **E** is based on Maxwell's equations. For microwave devices, these equations are combined to produce the Helmholtz vector wave equation

$$\nabla \times \nabla \times \mathbf{E} - k^2\mathbf{E} = 0$$

where k is the complex-valued wavenumber given by $k^2 = \omega^2\mu\varepsilon - j\omega\mu\sigma$ and μ is the magnetic permeability of the medium, which for tissue is the same as the free-space value, and ε is the electrical permittivity of the medium. The divergence free condition, $\nabla \cdot \mathbf{E} = 0$, may have to also be explicitly imposed if the solution technique does not inherently do this.

For RF devices, the frequency is sufficiently low that the displacement currents can be ignored. In this case, it is usually simpler to determine the scalar electric potential V and from this derive the electric field, $\mathbf{E} = -\nabla V$. The electric potential obeys a Poisson-type equation

$$-\nabla \cdot (k\nabla V) = 0$$

For models of both microwave and RF devices, the governing Helmholtz or Poisson equation is imposed in a domain with a known electric field or electric potential specified as a boundary condition to represent the power source. Another condition that is often imposed on the surface of metals is that the tangential component of the electric field is zero, $\hat{n} \times \mathbf{E} = 0$.

The solution of the governing equations with appropriate boundary conditions is impossible for all but the simplest geometries. For most practical cases, numerical methods and computational tools are required. The finite difference time domain (FDTD) method (7), the finite element (FE) method (8,9), and the volume surface integral equation (VSIE) method (10) are the most commonly utilized methods for solving the governing equations in electromagnetic hyperthermia and thermal therapy. In the FDTD method, the domain is discretized into rectangular elements. The accuracy of a FDTD solution depends on the size of the mesh spacing. Smaller elements produce more accurate solutions, but also require more memory to store the system of equations. Since the grids are rectangular, their nonconformation to curved tissue boundaries produces a stair-casing effect. Therefore, a large number of elements are required to model such geometries accurately. Unlike the FDTD method, the FE method uses tetrahedral meshes in the domain and the VSIE method uses triangular meshes on domain surfaces. Tetrahedral and triangular meshes are more suitable than regular finite difference grids for three-dimensional (3D) modeling since they do not have the stair casing effect at tissue boundaries.

RADIO FREQUENCY DEVICES

In RF, thermal therapy tissue is heated by electrical resistive (or J) heating. The heating devices, or applicators, are inserted interstitially to produce currents in the tissue. The currents typically oscillate sinusoidally in the kilohertz or low megahertz frequency range. As a result, this modality is often referred to as radio frequency or RF heating. There devices have an advantage over other interstitial devices in their simplicity and low cost. They can operate at low frequency, and therefore do not require complex power generators. The RF probes, due their simplicity, tend to have the smallest diameter of all the types of interstitial heating probes. The RF heating technique has been extensively reviewed by others (11–15).

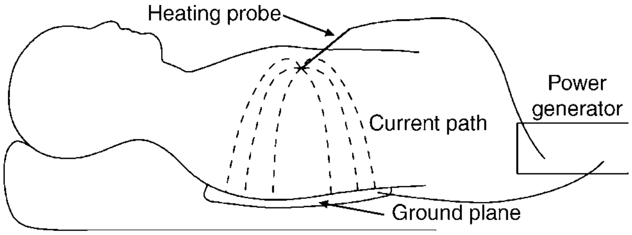


Figure 1. The heating in RF devices is caused by current flow. Since the current flows from the heating electrode to the ground pad, there is a high current density near the electrode due to its small size compared to the ground pad. This results in heating that is localized to the heating electrode.

There are several designs of RF interstitial devices, which may be categorized into three groups. The simplest design consists of a single electrode at a probe tip (often referred to as a needle electrode) (9,16–20). The current flows between a single electrode at the end of an applicator and a large ground plate placed at a distal site. Since the current flows between a small electrode and a large plate, the currents are concentrated near the electrodes resulting in SAR patterns that are localized to the electrodes as illustrated in Fig. 1.

With these single electrode probes the coagulation diameter is usually limited to ~ 1.6 cm. Therefore several probes are needed to cover a larger area (21), or a single probe can be inserted into several locations, sequentially, during a treatment.

Since it is desirable to avoid the insertion of multiple interstitial probes, single probes that release multiple electrodes outward from the probe tip have been designed to produce large coagulation volumes. Two examples of these are the Boston Scientific (Watertown, MA; formerly Radio Therapeutics Corporation, Mountain View, CA) RF 3000 system in which 10–12 tines are deployed from a cannula to form an umbrella shape (Fig. 2) and the RITA Medical Systems (Mountain View, CA) Starburst probes with up to 9 tines. In some configurations, some of the tines in the Starburst probes are replaced with dedicated thermocouples while others are hollow electrodes through which saline can be infused into the target region to enhance heating. These multielectrode probes are able to produce coagulation regions with diameters up to 7 cm, although complete coverage of a large region can be difficult in high blood flow organs, such as the kidney (22).

The negative RTOG phase III trial, in which only 1 out of 86 patients was deemed to have received an adequate thermal dose (1) illustrated the need to not only increase the target volume coverage, but also to control the heating.

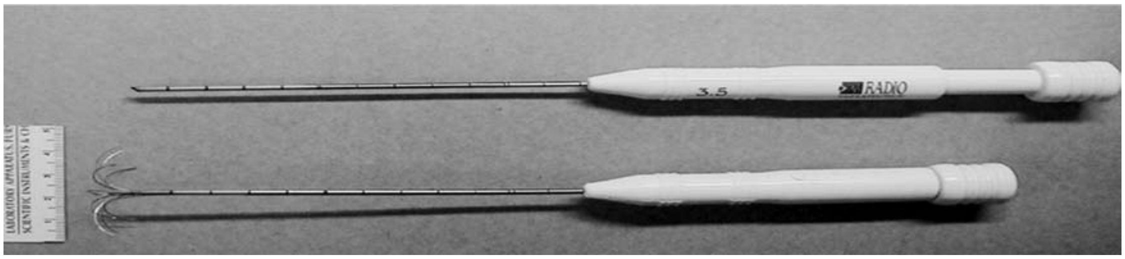


Figure 2. A Boston Scientific (Watertown, MA; formerly Radio Therapeutics Corporation, Mountain View, CA) insterstital RF probe with 10 tines that are deployed from the cannulus after insertion into a target region. The deployed tines produce a coagulation zone that is larger than the zone that can be produced by a single electrode probe. The top probe is shown with the tines undeployed (ready for insertion) and the bottom probe shows the probe with the tines deployed (as they would be after insertion).

Control is needed to enable the conformation of the heating to irregularly shaped target volumes while avoiding nearby organs at risk and to compensate for heterogeneous cooling by the vasculature (23). Partial control can be achieved by appropriate positioning of the probes and the adjustment their power. Further control along the direction of the probe is also needed (24,25) and multielectrode current source (MECS) applicators have been developed to provide this capability (26). The MECS applicators contain several electrodes placed along their length with the amplitude and phase of each electrode independently controlled. In the most common configuration, the electrodes are capacitively coupled (insulated) with the tissue. The electric fields induced by the electrodes produce currents in the tissue that cause heating. Since the electrodes are capacitively coupled, the probes can be inserted into brachytherapy catheters, for example, making it feasible to add interstitial heating as a simultaneous adjuvant to brachytherapy (interstitial radiation therapy). The electric field (and hence current) may be induced between electrodes on the same probe or on separate probes, or it may be induced between the probe electrodes and a grounding plane.

MICROWAVE DEVICES

Microwave applicators can produce larger coagulation regions than RF applicators due to their radiative nature. However, the construction of the power generator and matching circuitry makes these devices more complex, and therefore more expensive. Due to this, microwave interstitial hyperthermia has been used less often in the clinic than RF interstitial hyperthermia.

Ryan et al. reviewed and compared several types of microwave interstitial applicators (27) and several excellent reviews of microwave interstitial thermal therapy exist (28–32). The two most commonly used devices are the dipole antenna and the helical antenna. The dipole antenna is the simplest form of microwave interstitial antenna (7,8,33). It is usually constructed from a coaxial cable with the outer conductor removed from an end section (typically 1 or 2 cm in length) to expose the inner conductor (Fig. 3). A power generator feeds a sinusoidally oscillating signal into the cable at one of the ISM frequency bands between 400 MHz and 3 GHz. The inner- and outer-conductor electrodes at the tip of the coaxial cable act as an antenna that produces microwaves that radiate out into the tissue. Tissue is an attenuating medium that absorbs microwaves, and this absorbed energy is converted into heat in the tissue.

The radiative or active length of a typical dipole interstitial device is 1–3 cm. The devices produce a coagulation region that is ellipsoidal shaped with a large axis of up to 3 cm along the length of the antenna and a small axis of up to 2 cm diameter. The drawback of the dipole applicator is that the region of highest SAR, or hot spot, is located at the point at which the outer conductor is cut away. Therefore, the tips of these antennas have to be inserted past the center of the target region, and this can be a problem if the target region is located adjacent to a critical structure.

A further problem with dipole antennas is that the SAR patterns are sensitive to the depth to which the antenna is inserted into tissue (8). A second common microwave applicator design, referred to as a helical antenna (34–36), has been designed to make the applicator insensitive to its insertion depth. In this applicator, one electrode is wrapped in a helix pattern around an exposed coaxial cable (Fig. 4). The antennas are also designed to extend the heating pattern along the applicator and toward the tip of the antenna compared to the dipole antenna. The SAR pattern from a BSD Medical (Salt Lake City, UT) helical antenna is shown in (Fig. 5). The antenna was operating at 915 MHz. The measurement was performed using the thermographic imaging technique (37) and demonstrates

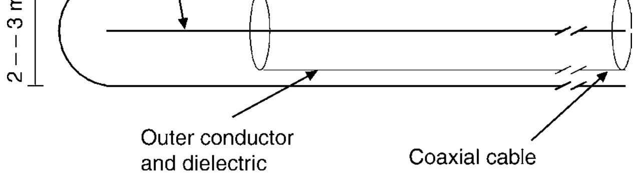


Figure 3. A schematic representation of a microwave interstitial dipole antenna applicator. The outer conductor of a coaxial cable is stripped away to produce a radiating section.

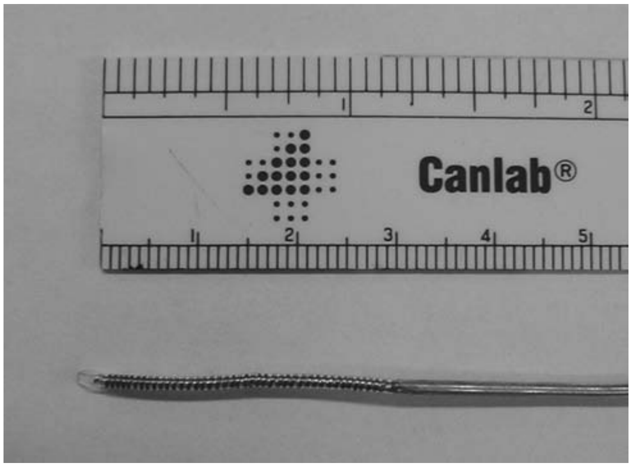


Figure 4. Shown here is a BSD Medical (Salt Lake City, UT) helical microwave applicator. The inner conductor of a coaxial cable is extended backward in a helical pattern around the dielectric insulator. There is no connection between the helical section and the outer conductor.

that the heating extends along the length of the helix and that the hot spot is close to the tip of the applicator.

Interstitial microwave applicators have the advantage over RF applicators in the ability to use arrays of applicators to dynamically steer the SAR pattern (33). For large target volumes, several applicators can be inserted. The heating pattern can then be adjusted by not only adjusting the power to each applicator, but also by adjusting the relative phase of the signal to each applicator. The phase can be adjusted such that the microwaves produced by the applicators interfere constructively in regions that require heating and interfere destructively in regions that should be spared. The predetermination of the phase required for each applicator can be calculated during treatment planning. This is a challenging calculation for applications in which tissue is electrically heterogeneous or the placement of the applicators cannot be accurately predicted. In these cases real-time monitoring of the treatment is required and a manual or computer run feedback control is used to set the phase of the applicators to produce the desired heating profile.

The size of the coagulation volume is limited by the maximum temperature in the treatment field. Since the maximum temperature is usually located at the applicator, it is possible to increase the coagulation volume by cooling adjacent to the applicator. Using this technique, the cross-section area of a coagulation volume has been noted to increase by a factor of 2.5 in one study (38) and the coagulation volume diameter was found to increase from 1.2 to 2.4 cm (39). In microwave heating, the cooling is usually done by passing water or air through the catheter containing the antenna (29,38,40). In RF heating, cooling water is passed inside the electrode to cool the tissue near the electrode (41,42).

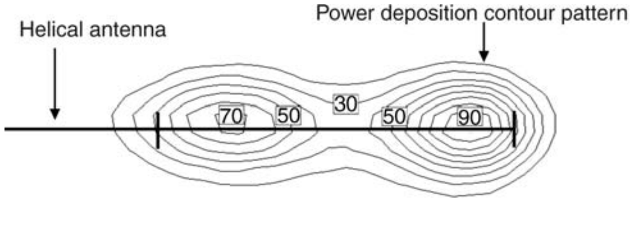


Figure 5. The normalized SAR pattern along the coronal plane of a BSD Medical (Salt Lake City, UT) helical applicator operating at 915 MHz. The image was provided courtesy of Claire McCann.

In RF heating, it is also possible to increase the coagulation volume by saline injection from the lumen of the electrode (43). Since saline is electrically conductive, injecting it into the tumor increases the electrical conductivity of the tumor, and hence the SAR in the tumor. This technique has not gained popularity due to the inability of control the flow of saline in the tumor, resulting in irregular and unpredictable coagulation regions being produced.

CLINICAL STUDIES WITH MICROWAVE AND RF DEVICES

Interstitial microwave and RF heating systems have been widely used to clinically treat a variety of tumors in phase I (34,44–47), phase II (18,32,44–49) and phase III trials (1,50). The RF systems have been used to treat a large range of sites, including brain (45), head and neck (1), breast (1), myocardium (51), lung (14), liver (11,52), pancreas (18), prostate (48), and kidney (44,53). Microwave systems have also been used to treat a large range of sites, including liver (4), prostate (both carcinoma and benign hyperplasia) (29,36), head and neck (1,32,49), brain (34,50), breast (1), and other pelvic areas (1). The heat treatments are used alone (29,36), or combined with external beam radiation (54), interstitial radiotherapy (brachytherapy) (1,32,46), and/or chemotherapy (17). The heat treatments are used alone (29,36), or combined with external beam radiation (29,36,48,54,55), combined with external beam radiation (54) or interstitial radiotherapy (brachytherapy) (1,32,46,48,55), and with chemotherapy (17).

The interstitial hyperthermia treatments are usually administered under ultrasound, CT or MR guidance. During the treatment the hyperechoginicity of microbubbles that can be produced at sufficiently high temperatures can provide some real-time ultrasound feedback of the treatment. Posttreatment evaluation can be performed using contrast enhanced ultrasound, CT or MR. The vasculature in the coagulated volume is destroyed and the destroyed volume can be identified as an unenhanced region in the image (41,56,57).

LASER DEVICES

First described in 1983 by Bown (58), Interstitial Laser Photocoagulation (ILP) [sometimes referred to as Laser Induced Thermal Therapy (LITT)] involves the use visible or near infra-red (IR) light delivered through fibre optic cables to heat tissue for therapeutic purposes. The ILP has been investigated as an experimental treatment for a variety of solid tumors including liver, breast, stomach, pancreas, kidney, lung, and bone (59). The tissue temperature is raised causing coagulation of the target volume. Similar to the microwave and RF cases, the production of heat in a local volume of tissue results from the amount of absorbed laser energy, $S(r)$. In biomedical treatments, such as LITT, it is the total absorbed laser energy that typically determines the therapeutic outcome. It is equal to the product of the local fluence rate, $\phi(r)$ which is the total photon power over all directions that pass through a point area of space, and the absorbing characteristics, $\mu_a(r)$ of the tissue (60):

$$S(r) = \mu_a(r)\phi(r)$$

The absorbed optical energy deposition pattern is governed by the absorption and scattering characteristics of the tissue. An absorption event causes the interacting molecule to enter a vibrational–rotational state that results in a transfer of energy to surrounding molecules that manifests as a local increase in temperature (61). Absorption occurs due to interactions with native molecules called chromophores with examples including melanin, hemoglobin, and water. In a given tissue, the concentration weighted sum of the absorption of different chromophores leads to its bulk macroscopic absorption. Scattering refers to a directional change in light propagation and likely results from differences in the index of refraction in the various cellular components, such as the between cell membranes and the extracellular space. Here the scattering is assumed to be elastic with no change in energy occurring during the interaction. The statistical quantities that govern light interactions are the scattering coefficient, $\mu_a(\text{cm}^{-1})$ and absorption coefficient, $\mu_a(\text{cm}^{-1})$ and are defined, respectively, as the probability of scattering or absorption per average distance traveled (also known as the mean free path). In the case of scattering, consideration is given to the probability of scatter in a particular direction. An additional parameter known as the anisotropy factor, g, quantifies this directionality by integrating the average cosine of the scattering probability over all directions. When $g = 0$, scattering is isotropic. However, in the case of biological tissues g typically lies within the range of 0.7 and 0.99 meaning that scattering typically occurs in the forward direction. The reduced scattering coefficient, $\mu_s' = \mu_s(1-g)$, allows light scattering to be approximated as isotropic although scattering events are actually in the forward direction. The inverse of the reduced scattering coefficient is, therefore, the average distance that light travels before it changes direction from its original direction of propagation (62).

In theory, Maxwell's equations could be used to calculate the scattering and absorption of the EM vector fields due to the underlying tissue components (63). In this case, the tissue microstructure could be modeled as random perturbations, $\varepsilon_1(r)$ in the dielectric constant around a mean value, $\varepsilon_0(r)$, with the total dielectric constant, $\varepsilon(r)$, given by the sum of these quantities. However, in practice, due to the complex and random composition of tissue, a complete and accurate description of $\varepsilon(r)$ has yet to be realized. Instead a more commonly used solution is to consider light as a stream of neutral particles or photons with individual quanta of energy that propagate elastically throughout the medium. This formalism is governed by radiative transport theory (64), and assumes light to be monochromatic while ignoring its conventional wave characteristics, such as polarization, diffraction, interference, and fluorescence. Although incomplete, the photon model has been shown to be consistent with experimental measurements in turbid media (65).

A commonly employed model of photon propagation is the Monte Carlo (MC) method (66), which utilizes probability distributions to simulate the propagation of thousands to millions of individual photon packets based on the optical properties of tissue to arrive at a statistical representation of the overall light distribution. The MC is amenable to heterogeneous and arbitrary geometries and does not suffer from the limiting assumptions of analytical solutions. However, its primary disadvantage is the requirement of long computational times, on the order of hours to days, to achieve reasonable statistics. Regardless, with the increasing speed of modern computers, the Monte Carlo method remains a viable option for photon simulations. The reader is referred to an excellent review by Roggan and Muller (67) for the implementation of the MC model for treatment planning of LITT.

Alternatively, one may employ formal solutions to the governing equations for photon transport. The energy flow of photons in a scattering and absorbing medium is described by the radiative transfer equation (RTE) (64). The RTE is an integro differential equation that describes the energy conservation of photons within an infinitesimally small volume that result from losses due to absorption and scattering as well as gains arising from photons scattered from other directions and from the laser source. Analytical solutions to the RTE are difficult to obtain. Hence, various approximations have been proposed to convert the RTE to a more mathematically tractable and practical form. A standard technique, called the P_n approximation, expands the radiance and source as a finite series of spherical harmonics to nth order. The $P1$ approximation is the simplest of these expansions and in the steady state is also known as the diffusion approximation (63,64):

$$\nabla^2\phi(\vec{r}) - \frac{\mu_a}{D}(\vec{r})\phi(\vec{r}) = -\frac{1}{D}S(\vec{r})$$

Here $\phi(\vec{r})$ is the photon fluence rate, while D is the photon diffusion coefficient given by

$$D = \frac{1}{3[\mu_\text{s}' + \mu_\text{a}]}$$

The primary assumption of the diffusion equation, that is linear flux anisotropy, is only accurate when the scattering properties of the medium are much larger than the absorption properties and at locations $> 1/\mu_\text{s}'$ from the source. A number of analytical solutions to the diffusion equation exist for simple but practical geometries. The solution for a point source in an infinite homogeneous medium is given by (63)

$$\phi(\vec{r}) = \frac{P_\text{o}e^{(-\mu_\text{eff}r)}}{4\pi r}$$

This solution is particularly useful as, assuming an infinite medium, it may be integrated numerically to provide the light distribution of cylindrical or extended source of arbitrary geometries. However, it is well known that tissue optical properties often change from their native state after undergoing thermal coagulation. This results in heterogeneities in optical properties that effect the overall light distribution (68). In such cases, analytical solutions are available only for the simplest geometries and numerical methods such as the finite element (69), finite difference (70), and boundary element method (71) must be employed. A thorough discussion of these methods was given in the preceding section for microwaves and their implementation in the case of photon propagation is the same.

Initially, bare tipped optical fibers were used to deliver laser light to the tumor. High temperatures immediately adjacent to the fiber tip cause the tissue to char and form a zone of carbonization. The charred fiber then acts as a point heat source and the temperature of the fiber increases significantly leading to vacuolization of the surrounding tissue. The volume of coagulation around the fiber grows until thermal equilibrium is reached at the edges of the lesion. Here, the conduction of heat from the fiber is balanced by the tissue's ability to remove energy through blood flow and thermal conduction.

The size of the lesion depends on the thermal conduction properties of the tissue, but would normally be limited to ~ 2 cm in diameter. Larger tumors require multiple optical fiber implants to enable complete coverage of the tumor volume. For example, a 4 cm diameter tumor would require at least eight fibers to fully coagulate the tumor.

The limitations of the bare tipped fibers have been addressed in two ways. The first was to employ a line source geometry instead of a point source. This can be achieved by using a diffusing tip fiber where light gradually leaks out of the fiber over an extended distance of a few centimeters. The second approach is to restrict the temperature of the fiber to lower than the charring threshold by controlling the power delivered to the fiber. If charring is avoided, light can propagate into the tissue resulting in heating at a distance from the fiber and a broader SAR pattern. These two approaches can be combined to achieve greater lesion volumes from single fibers. Heisterkamp et al. (72) demonstrated an almost doubling of the coagulated volume from 4.32 cm^3 (bare tipped) to 8.16 cm^3 (temperature restricted diffusing tip) using such an approach.

The other major factor that affects the lesion size is the wavelength of the light used. Somewhat counterintuitively, light that is less absorbed by tissue, results in greater lesion sizes. This is because the light can penetrate further into the tissue, and therefore directly heat at greater distances from the fiber. The availability of high power sources at two specific wavelengths (810 nm as produced by diode lasers and 1064 nm as produced by Nd:YAG lasers) has dominated the development of interstitial laser thermal therapy. Wyman et al. (73) have shown that 1064 nm light can enable the creation of greater lesion sizes due to its greater penetration. However, Nd:YAG lasers are large, generally immobile and inconvenient and so many have adopted 810 nm as the wavelength of choice due to the availability of compact and inexpensive sources. More recently 980 nm lasers have been employed to combine mobility with greater light penetration (74,75).

Differences between Nd:YAG and Diode lasers are only realized if charring is avoided. Once charring and carbonization has occurred the fiber acts as a point or line heat source. There is no further light propagation into the tissue and subsequent heating has no wavelength dependency. In order to exploit the penetration of light into the tissue, the fiber tip temperature must be controlled to avoid charring. Achieving such control is somewhat challenging as the temperature of the tip can rise very quickly in a positive feedback loop. As charring begins, the rate of temperature rise increases that causes an increasing rate of charring. Robust, automatic feedback control mechanisms are necessary to ensure controlled heating and lesion formation.

INTERSTITIAL ULTRASOUND

The possibility of developing interstitial ultrasound devices for hyperthermia applications was proposed by Hynynen in 1992 (76). The initial studies examined various design parameters including the choice of ultrasound frequency, electric and acoustic power, and catheter cooling. As Hynynen showed (76), thin interstitial ultrasound applicators were likely capable of heating perfused tissue to therapeutic temperatures.

Ultrasound is a high frequency longitudinal pressure wave that can pass relatively easily through soft tissue. Consequently, it has been useful as an energy source for diagnostic imaging where focussed ultrasound radiators are used to produce high resolution images of soft tissue abnormalities. During transmission through tissue energy is lost due to absorption and to a much lesser extent to scattering. The absorption is caused by friction as the pressure wave causes relative motion of the tissue components. These frictional forces cause heating that can be significant if the incident ultrasound power is high enough. The absorption, α is frequency dependent where

$$\alpha = a f^m$$

and a and m are coefficients that are variable between tissues although m is ~ 1.5 for most soft tissues. Rapidly increasing absorption with frequency is the main reason that the penetration of diagnostic imaging is limited at very high ultrasound frequencies. Higher penetration is also the reason that relatively low ultrasound frequencies are used for ultrasound heating. Typically, frequencies in the range 0.5–2 MHz have been used in external focused ultrasound heating applications. However, this becomes problematic for interstitial devices that are small and resonate at high ultrasound frequencies.

Interstitial ultrasound applicators have since been developed and are usually designed as thin tubular radiators. The radiator consists of a piezoelectric material that will resonate acoustically at a frequency f determined by the wall diameter d:

$$f = \frac{v}{2d}$$

where v is the speed of sound in the piezoelectric material (e. g., 4000 m·s^{-1} in the piezoelectric material PZT 4A). For interstitial applicators, thin radiators are required. A wall thickness of 0.2 mm, for example, would translate into an operating frequency of ~ 10 MHz (76). The SAR for a cylindrical applicator is dependent on its dimensions and the frequency of operation as given by

$$\mathrm{SAR} = 2\alpha\, f I_0 \left(\frac{r}{r_0}\right) \mathrm{e}^{-2\mu\, f(r-r_0)}$$

where α is the ultrasound absorption coefficient in tissue, I_0 is the intensity of ultrasound at the applicator surface, r_0 is the radius of the applicator, r is the distance from the centre of the applicator to the point of interest and μ is the attenuation coefficient of ultrasound that includes absorption and scattering. Skinner et al. (77) have calculated and compared the SAR of ultrasound, laser, and microwave applicators assuming a simple cyclindrical radiation pattern for each. The SAR of all these applicators is dominated by the thin cylindrical geometry so that despite the larger penetration depth of ultrasound, only slightly larger diameter lesions can be produced. In order to overcome the limiting geometry, new interstitial ultrasound applicators have been developed that take advantage of the focusing ability of ultrasound (78) or that employs acoustic matching that can result in efficient transmission at multiple frequencies (79).

The development of interstitial ultrasound applicators is still at the preclinical stage (80,81) although larger, intracavitary applicators are being applied in the treatment of prostate cancer using a transrectal technique (82).

BIBLIOGRAPHY

1. Emami BC, et al. Phase III study of interstitial thermoradiotherapy compared with interstitial radiotherapy alone in the treatment of recurrent or persistent human tumors: A prospectively controlled randomized study by the Radiation Therapy Oncology Group. Inte J Rad Oncol Biol Phy 1996;34(5): 1097–1104.
2. Emami BP, et al. RTOG Quality Assurance Guidelines for Interstitial Hyperthermia. Inter J Rad Oncol Biol Phy 1991; 20(5):1117–1124.
3. Peters RD, et al. Magnetic resonance thermometry for predicting thermal damage: An application of interstitial laser coagulation in an in vivo canine prostate model. Magn Reson Med 2000;44(6):873–883.
4. Morikawa S, et al. MR-guided microwave thermocoagulation therapy of liver tumors: Initial clinical experiences using a 0.5 T open MR system. J Magn Reson Imaging 2002;16(5):576–583.
5. Sapareto SA, Dewey WC. Thermal dose determination in cancer therapy. Int J Radiat Oncol Biol Phys 1984;10(6): 787–800.
6. Pennes HH .Analysis of tissue and arterial blood temperatures in the resting human forearm. 1948. J Appl Physiol 1998; 85(1):5–34.
7. Camart JC, et al. New 434 MHz interstitial hyperthermia system monitored by microwave radiometry: theoretical and experimental results. Intern J Hypertherm 2000;16(2):95–111.
8. Mechling JA, Strohbehn JW. 3-Dimensional Theoretical SAR and Temperature Distributions Created in Brain-Tissue by 915 and 2450 MHz Dipole Antenna-Arrays with Varying Insertion Depths. Intern J Hypertherm 1992;8(4):529–542.
9. Uzuka T, et al. Planning of hyperthermic treatment for malignant glioma using computer simulation. Int J Hypertherm 2001;17(2):114–122.
10. Wust P, et al. Simulation studies promote technological development of radiofrequency phased array hyperthermia. Int J Hypertherm 1996;12(4):477–494.
11. Haemmerich D, Lee Jr FT. Multiple applicator approaches for radiofrequency and microwave ablation. Int J Hypertherm 2005;21(2):93–106.
12. McGahan JP, Dodd GD, 3rd. Radiofrequency ablation of the liver: current status. AJR Am J Roentgenol 2001;176(1):3–16.
13. Friedman MI, et al. Radiofrequency ablation of cancer. Cardiovasc Intervent Radiol 2004;27(5):427–434.
14. Lencioni RL, et al. Radiofrequency ablation of lung malignancies: where do we stand? Cardiovasc Intervent Radiol 2004;27(6): 581–590.
15. Gazelle GS, Goldberg SN, Solbiati L, Livraghi T. Tumor ablation with radio-frequency energy. Radiology 2000;217(3): 633–646.
16. Goletti O, et al. Laparoscopic radiofrequency thermal ablation of hepatocarcinoma: preliminary experience. Surg Laparosc Endosc Percutan Tech 2000;10(5):284–290.
17. Morita K, et al. Combination therapy of rat brain tumours using localized interstitial hyperthermia and intra-arterial chemotherapy. Inter J Hypertherm 2003;19(2):204–212.
18. Matsui Y, et al. Selective thermocoagulation of unresectable pancreatic cancers by using radiofrequency capacitive heating. Pancreas 2000;20(1):14–20.
19. Aoki H, et al. Therapeutic efficacy of targeting chemotherapy using local hyperthermia and thermosensitive liposome: evaluation of drug distribution in a rat glioma model. Int J Hypertherm 2004;20(6):595–605.
20. Lencioni R, et al. Radio-frequency thermal ablation of liver metastases with a cooled-tip electrode needle: results of a pilot clinical trial. Eur Radiol 1998;8(7):1205–1211.
21. Haemmerich D, et al. Large-volume radiofrequency ablation of ex vivo bovine liver with multiple cooled cluster electrodes. Radiology 2005;234(2):563–568.
22. Rendon RA, et al. The uncertainty of radio frequency treatment of renal cell carcinoma: Findings at immediate and delayed nephrectomy. J Urol 2002;167(4):1587–1592.
23. Crezee J, Lagendijk JJ. Temperature uniformity during hyperthermia: the impact of large vessels. Phys Med Biol 1992;37(6):1321–1337.
24. vanderKoijk JF, et al. Dose uniformity in MECS interstitial hyperthermia: The impact of longitudinal control in model anatomies. Phys Med Biol 1996;41(3):429–444.
25. VanderKoijk JF, et al. The influence of vasculature on temperature distributions in MECS interstitial hyperthermia: Importance of longitudinal control. Intern J Hypertherm 1997; 13(4):365–385.
26. Kaatee RSJP. Development and evaluation of a 27 MHz multi-electrode current-source interstitial hyperthermia system. Med Phys 2000;27(12):2829–2829.
27. Ryan TP. Comparison of 6 Microwave Antennas for Hyperthermia Treatment of Cancer—SAR Results for Single Antennas and Arrays. Intern J Rad Oncol Biol Phys 1991; 21(2):403–413.
28. Roemer RB. Engineering aspects of hyperthermia therapy. Annu Rev Biomed Eng 1999;1:347–376.
29. Sherar MD, Trachtenberg J, Davidson SRH, Gertner MR. Interstitial microwave thermal therapy and its application to the treatment of recurrent prostate cancer. Intern J Hypertherm 2004;20(7):757–768.
30. Fabre JJ, et al. 915 MHz Microwave Interstitial Hyperthermia. 1. Theoretical and Experimental Aspects with Temperature Control by Multifrequency Radiometry. Intern J Hypertherm 1993;9(3):433–444.
31. Camart JC, et al. 915 MHz Microwave Interstitial Hyperthermia. 2. Array of Phase-Monitored Antennas. Intern J Hypertherm 1993;9(3):445–454.

32. Prevost B, et al. 915 MHz Microwave Interstitial Hyperthermia. 3. Phase-II Clinical-Results. Intern J Hypertherm 1993; 9(3):455–462.
33. Camart JC, et al. Coaxial Antenna-Array for 915 MHz Interstitial Hyperthermia—Design and Modelization Power Deposition and Heating Pattern Phased-Array. IEEE Trans Microwave Theory Tech 1992;40(12):2243–2250.
34. Fike JR, Gobbel GT, Satoh T, Stauffer PR. Normal Brain Response after Interstitial Microwave Hyperthermia. Intern J Hypertherm 1991;7(5): 795–808.
35. McCann C, et al. Feasibility of salvage interstitial microwave thermal therapy for prostate carcinoma following failed brachytherapy: studies in a tissue equivalent phantom. Phys Med Biol 2003;48(8):1041–1052.
36. Sherar MD, et al. Interstitial microwave thermal therapy for prostate cancer. J Endourol 2003;17(8):617–625.
37. Guy A. Analysis of Electromagnetic Fields Induced in Biological Tissues by Thermographic Studies on Equivalent Phanton Models. IEEE Trans Biomed Eng 1971;19:205–214.
38. Trembly BS, Douple EB, Hoopes PJ. The Effect of Air Cooling on the Radial Temperature Distribution of a Single Microwave Hyperthermia Antenna In vivo. Intern J Hypertherm 1991;7(2):343–354.
39. Goldberg SN, et al. Radiofrequency tissue ablation: increased lesion diameter with a perfusion electrode. Acad Radiol 1996;3(8):636–644.
40. Eppert V, Trembly BS, Richter HJ. Air Cooling for an Interstitial Microwave Hyperthermia Antenna—Theory and Experiment. IEEE Trans Biomed Eng 1991;38(5):450–460.
41. Goldberg SN, et al. Treatment of intrahepatic malignancy with radiofrequency ablation: Radiologic-pathologic correlation. Cancer 2000;88(11):2452–2463.
42. Solbiati L, et al. Hepatic metastases: percutaneous radiofrequency ablation with cooled-tip electrodes. Radiology 1997; 205(2):367–373.
43. Livraghi T, et al. Saline-enhanced radio-frequency tissue ablation in the treatment of liver metastases. Radiology 1997;202(1):205–210.
44. Michaels MJ, et al. Incomplete renal tumor destruction using radio frequency interstitial ablation. J Urol 2002;168(6):2406–2409.
45. Takahashi H, et al. Radiofrequency interstitial hyperthermia of malignant brain tumors: Development of heating system. Exper Oncol 2000;22(4):186–190.
46. Seegenschmiedt MH, et al. Clinical-Experience with Interstitial Thermoradiotherapy for Localized Implantable Pelvic Tumors. Am J Clin Oncol Cancer Clin Trials 1993;16(3): 210–222.
47. Seegenschmiedt MH, et al. Multivariate-Analysis of Prognostic Parameters Using Interstitial Thermoradiotherapy (Iht-Irt)—Tumor and Treatment Variables Predict Outcome. Intern J Rad Oncol Biol Phys 1994;29(5):1049–1063.
48. van Vulpen M, et al. Radiotherapy and hyperthermia in the treatment of patients with locally advanced prostate cancer: Preliminary results. Bju Inter 2004;93(1):36–41.
49. Engin K, et al. Thermoradiotherapy with Combined Interstitial and External Hyperthermia in Advanced Tumors in the Head and Neck with Depth Greater-Than-or-Equal-to 3 Cm. Intern J Hyperther 1993;9(5):645–654.
50. Sneed PK, et al. Thermoradiotherapy of Recurrent Malignant Brain-Tumors. Int J Radiat Oncol Biology Physics. 1992.
51. Wonnell TL, Stauffer PR, Langberg JJ. Evaluation of Microwave and Radio-Frequency Catheter Ablation in a Myocardium-Equivalent Phantom Model. IEEE Trans Biomed Eng 1992;39(10):1086–1095.
52. Buscarini L, Buscarini E. Therapy of HCC-radiofrequency ablation. Hepato-Gastroenterol 2001;48(37):15–19.
53. Rendon RA, et al. Development of a radiofrequency based thermal therapy technique in an in vivo porcine model for the treatment of small renal masses. J Urol 2001;166(1):292–298.
54. Blute ML, Larson T. Minimally invasive therapies for benign prostatic hyperplasia. Urology 2001;58(6A):33–40.
55. Van Vulpen M, et al. Three-dimensional controlled interstitial hyperthermia combined with radiotherapy for locally advanced prostate carcinoma—A feasibility study. Intern J Rad Oncol Biol Phy 2002;53(1):116–126.
56. Belfiore G, et al. CT-guided radiofrequency ablation: a potential complementary therapy for patients with unresectable primary lung cancer—a preliminary report of 33 patients. AJR Am J Roentgenol 2004;183(4):1003–1011.
57. Cioni D, Lencioni R, Bartolozzi C, Percutaneous ablation of liver malignancies: Imaging evaluation of treatment response. Eur J Ultrasound 2001;13(2):73–93.
58. Bown SG. Phototherapy in tumors. World J Surg 1983;7(6): 700–709.
59. Witt JD, et al. Interstitial laser photocoagulation for the treatment of osteoid osteoma. J Bone Joint Surg Br 2000; 82(8):1125–1128.
60. Welch A. The thermal response of laser irradiated tissue. IEEE J Quantum Electr 1984;20(12):1471–1481.
61. Boulnois J. Photophysical processes in recent medical laser developments: A review. Lasers Med Sci 1986;1(1):47–66.
62. Wyman D, Patterson M, Wilson B. Similarity relations for the interaction parameters in radiation transport and their applications. Appl Op 1989;28:5243–5249.
63. Ishimaru A. Diffusion Approximation, in Wave Propagation and Scattering in Random Media. New York: Academic Press; 1978. p. 178.
64. Duderstadt JH. Nuclear Reactor Analysis. New York: John Wiley & Sons; 1976.
65. Rinzema K, Murrer L. Direct experimental verification of light transport theory in an optical phantom. J Opt Soc Am A 1998;15(8):2078–2088.
66. Wilson BC, Adam G. A Monte Carlo model for the absorption and flux distributions of light in tissue. Med Phys 1983; 10(6):824–830.
67. Roggan A, Muller G. Dosimetry and computer based irradiation planning for laser-induced interstitial thermotherapy (LITT). In: Roggan A, Muller G, editors. Laser-Induced Interstitial Thermotherapy. Bellingham, (WA): SPIE Press; 114–156.
68. Jaywant S, et al. Temperature dependent changes in the optical absorptio nand scattering spectra of tissue. SPIE Proc 1882. 1993;
69. Arridge SR, Schweiger M, Hiraoka M, Delpy DT. A finite element approach for modeling photon transport in tissue. Med Phys 1993;20(2 Pt. 1):299–309.
70. Pogue BW, Patterson MS, Jiang H, Paulsen KD. Initial assessment of a simple system for frequency domain diffuse optical tomography. Phys Med Biol 1995;40(10):1709–1729.
71. Ripoll J, Nieto-Vesperinas M. Scattering Integral Equations for Diffusive Waves. Detection of Objects Buried in Diffusive Media in the Presence of Rough Interfaces. J Opt Soc of Am A 1999;16:1453–1465.
72. Heisterkamp J, van Hillegersberg R, Sinofsky E. Heat-resistant cylindrical diffuser for interstitial laser coagulation: Comparison with the bare-tip fiber in a porcine liver model. Lasers Surg Med 1997;20(3):304–309.
73. Wyman DR, Schatz SW, Maguire JA. Comparison of 810 nm and 1064 nm wavelengths for interstitial laser photocoagulation in rabbit brain. Lasers Surg Med 1997;21(1):50–58.
74. McNichols RJ, et al. MR thermometry-based feedback control of laser interstitial thermal therapy at 980 nm. Lasers Surg Med 2004;34(1):48–55.

75. Kangasniemi M, et al. Thermal therapy of canine cerebral tumors using a 980 nm diode laser with MR temperature-sensitive imaging feedback. Lasers Surg Med 2004;35(1):41–50.

76. Hynynen K. The Feasibility of Interstitial Ultrasound Hyperthermia. Med Phys 1992;19(4):979–987.

77. Skinner MG, Iizuka MN, Kolios MC, Sherar MD. A theoretical comparison of energy sources—microwave, ultrasound and laser—for interstitial thermal therapy. Phys Med Biol 1998; 43(12):3535–3547.

78. Chopra R, Bronskill MJ, Foster FS. Feasibility of linear arrays for interstitial ultrasound thermal therapy. Med Phys 2000; 27(6):1281–1286.

79. Chopra R, Luginbuhl C, Foster FS, Bronskill MJ. Multifrequency ultrasound transducers for conformal interstitial thermal therapy. IEEE Trans Ultrason Ferroelectr Freq Control 2003;50(7):881–889.

80. Nau WH, et al. MRI-guided interstitial ultrasound thermal therapy of the prostate: A feasibility study in the canine model. Med Phys 2005;32(3):733–743.

81. Diederich CJ, et al. Catheter-based ultrasound applicators for selective thermal ablation: Progress towards MRI-guided applications in prostate. Int J Hyperther 2004;20(7):739–756.

82. Uchida T, et al. Transrectal high-intensity focused ultrasound for treatment of patients with stage T1b-2NOMO localized prostate cancer: A preliminary report. Urology 2002; 59(3): 394–398.

See also BRACHYTHERAPY, HIGH DOSAGE RATE; HEAT AND COLD, THERAPEUTIC; HYPERTHERMIA, SYSTEMIC; HYPERTHERMIA, ULTRASONIC; PROSTATE SEED IMPLANTS.

HYPERTHERMIA, SYSTEMIC

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

INTRODUCTION

Systemic hyperthermia is deliberate heating of the whole body to achieve an elevated core temperature for therapeutic purposes. Other terms used are whole-body hyperthermia, systemic or whole body thermal therapy, and hyperpyrexia. The goal of systemic hyperthermia is to reproduce the beneficial effects of fever. Typically, core body temperatures of 41–42 °C are induced for 1–2 h, or alternatively 39–40 °C for 4–8 h. Systemic hyperthermia, by virtue of application to the whole body, aims to alleviate systemic disease conditions, in contrast to local or regional hyperthermia that treats only a specific tissue, limb, or body region.

HISTORICAL BACKGROUND

The use of heat to treat disease goes back to ancient times. Application of fire to cure a breast tumor is recorded in an ancient Egyptian papyrus, and the therapeutic value of elevated body temperature in the form of fever was appreciated by ancient Greek physicians. Hippocrates wrote, “What medicines do not heal, the lance will; what the lance does not heal, fire will,” while Parmenides stated, “Give me a chance to create a fever and I will cure any disease.” In the first century AD, Rufus (also written as Refus or Ruphos) of Ephesus advocated fever therapy for a variety of diseases. Hot baths were considered therapeutic in ancient Egypt, Greece, Rome, China, and India as they still are in many aboriginal cultures today, along with burying diseased individuals in hot sand or mud. Hot baths and saunas are an integral part of health traditions throughout the Orient, in Indian Ayurvedic medicine, as well as in Eastern European and Scandinavian countries. Following several earlier anecdotal reports, several nineteenth century German physicians observed regression or cure of sarcoma in patients who suffered prolonged, high fevers due to infectious diseases. This led to efforts to induce infectious fevers in cancer patients, for example, by applying soiled bandages or the blood of malaria patients to wounds. The late nineteenth century New York physician, William Coley, achieved cancer cures by administration of erysipelas and other bacterial endotoxins, now known as Coley’s toxins, and attempted to create standardized preparations of these pyrogens (1). At around the same time, treatment of syphilis by placing the patient in a stove-heated room, or a heat box, became commonplace. Successful hyperthermic treatment of other sexually transmitted diseases, such as gonorrhea, and neurological conditions, such as chorea minor, dementia paralytica, and multiple sclerosis along with arthritis, and asthma were widely reported. Interestingly, it was noted by Italian physicians that upon completion of the draining of the Pontine Swamps near Rome by Mussolini in the 1930s, not only was malaria eradicated, but the prevalence of cancer in the area was the same as in the rest of Italy, whereas earlier the whole malaria-infected region was noted for its absence of cancer. It was concluded that the frequent fever attacks common in malaria stimulated the immune system to prevent the development of cancers.

The science of hyperthermia became grounded in the first few decades of the twentieth century when some of the biological effects of elevated body temperature were elucidated and attempts were made to understand and control the therapeutic application of heat. Numerous devices were developed to produce elevated temperatures of the body, by a variety of physical means. After a shift in focus to local and regional hyperthermia, there is now a resurgence of interest in systemic hyperthermia for treatment of cancer, as well as other systemic diseases. Whole-body hyperthermia treatment is now carried out at several university centers in the United States, and Europe (Table 1), where controlled clinical trials are being carried out. Numerous private clinics, principally in North America, Germany, Austria, Eastern Europe, Japan, and China also perform systemic hyperthermia, mostly as part of holistic, alternative, treatment regimens.

PHYSICS OF SYSTEMIC HYPERTHERMIA

As shown schematically in Fig. 1, in order to achieve body temperature elevation, there must be greater deposition of heat energy in the body than heat energy lost from

Table 1. Clinical Academic/Regional Systemic Hyperthermia Centers

Country	City	Institution	Principal Investigator	Heat Type, Machine[a]	Protocol (time, temp)
Asia					
China	Baoding	Second Hospital of Baoding	Chunzhu Yin	RF	3.5–6 h, 40–40.5 °C
China	Changchun	Jilin Tumor Hospital	Changguo Hong	RF	3.5–6 h, 40–40.5 °C
China	Jiangmen	Guangdong Jiangmen Renmin Hospital	Wenping Wu	IR	
China	Shanghai	Changhai Hospital, Second Military Medical School	Yajie Wang		
China	Shanghai	Department of Tumor Hyperthermia Center	Kai-sheng Hou	IR, ET-Space extracorporeal	1–2 h, 41.8–42.5 °C
China	Shanghai	Shanghai Jingan Central Hospital	Weiping Tao	IR, ET-Space extracorporeal	2 h, 41.6 °C 4 h, 42.1 °C
China	Tai'an City	88th Hospital of PLA	Yong Peng		1–2 h, 41.8 °C 6 h, 39.5–40 °C
China	Zhengzhou	Modern Hospital, Zhengzhou	Dingjiu Li	RF	3.5–6 h, 40–40.5 °C
Japan	Tokyo	Luke Hospital	Akira Takeuchi	IR	
Europe					
Belarus	Minsk	Belarus Center for Pediatric Oncology and Hematology	Reimann Ismail-zade	HF EM, Yakhta-5	2 h, 41.8–42.5 °C 1 h, 42.5–43 °C
Germany	Berlin	Ludwig Maximilian University Charité Medical Center	Bert Hildebrandt, Hanno Riess, Peter Wust	IR, Iratherm	1 h, 41.8 °C
Germany	Frankfurt	Krankenhaus Nordwest	Elke Jäger, Akin Atmata	IR, Aquatherm	1 h, 41.8 °C
Germany	Munich	Ludwig Maximilian University Hospital Clinic	Harald Sommer	IR, Iratherm	1 h, 41.8 °C
Hungary	Kecskemét	Institute of Radiology of Kecskemét	Miklós Szücs	IR, OncoTherm	1 h, 41.8 °C
Norway	Bergen	University of Bergen, Haukeland University Hospital	Baard-Christian Schem	IR, Iratherm	1 h, 41.8 °C
Russia	Novosibirsk	Siberian Scientific Research Institute of Hyperthermia	Roman Tchervov	Water bath	43.5–44.0 °C
Russia	Obninsk	Medical Radiological Research Center of Russian Academy of Medical Sciences, Obninsk	Yuri Mardynsky	HF EM, Yakhta 5	1–2 h, 41.0–42.3 °C
North America					
United States	Galveston, TX	University of Texas Medical Branch	Joseph Zwischenberger	extracorporeal	2 h, 42.5 °C
United States	Houston, TX	University of Texas Medical School	Joan M. Bull	IR, Heckel	6 h, 40 °C
United States	Buffalo, NY	Roswell Park Cancer Institute	William G. Kraybill	IR, Heckel	6 h, 40 °C
United States	Durham, NC	Duke Comprehensive Cancer Center[b]	Zeljko Vujaskovic	IR, Heckel	6 h, 40 °C

[a] Radio frequency = RF; infrared = IR.
[b] Starting in 2006.

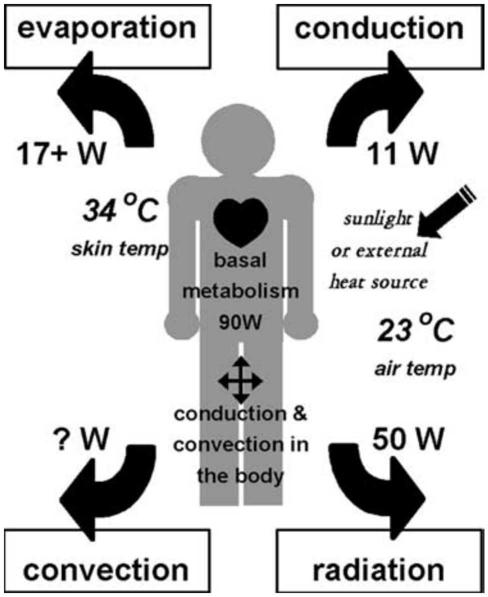


Figure 1. Schematic of heat balance mechanisms in the human body. Body temperature is determined by the balance of metabolic heat production plus heating from external sources, and heat losses by radiation, evaporation, convection, and conduction.

conduction, convection, radiation and evaporation, that is,

$$Q'_{\text{dep}}\Delta t > Q'_{\text{loss}}\Delta t \tag{1}$$

where $Q' = \Delta Q/\Delta t$ represents the change in heat energy, Q (measured in Joules or calories), over a time period Δt. Net heat energy deposition in a volume element ΔV of tissue of density ρ_{tis} results in an increase in temperature ΔT dependent on the specific heat of the tissue, c_{tis},

$$\left(\frac{Q'_{\text{dep}}}{\Delta V} - \frac{Q'_{\text{loss}}}{\Delta V}\right)\Delta t = (\rho_{\text{tis}}\Delta V)c_{\text{tis}}\Delta T$$

$$\Delta T = (Q'_{\text{dep}} - Q'_{\text{loss}}) \cdot \frac{\Delta t}{\rho_{\text{tis}}c_{\text{tis}}} \tag{2}$$

Heat deposition is the sum of the absorbed power density, P_{abs}, from external sources and heat generated by metabolism, Q_{met},

$$\frac{\Delta Q'_{\text{dep}}}{\Delta V} = P_{\text{abs}}(x,y,z,t) + \frac{\Delta Q'_{\text{met}}}{\Delta V}(x,y,z,t) \tag{3}$$

If the air temperature is higher than the body surface temperature, heat is absorbed from air surrounding the body by the skin, as well as during respiration. Power deposition in tissue from external electromagnetic fields depends on the coupling of the radiation field (microwave, RF, ultrasound, visible or IR light) with tissue. The body's metabolic rate depends on the amount of muscular activity, the temperature, pressure and humidity of the environment, and the size of the body. Metabolic rate increases nonlinearly with core body temperature, in part due to the exponential increase of the rate of chemical reactions with temperature (Arrhenius equation). An empirical relationship between basal metabolic rate and core temperature has been determined as

$$\text{Basal MR} = \frac{85 \times 1.07^{(T_{\text{core}})}}{0.5} \tag{4}$$

which can be exploited to maintain elevated body temperatures (2). At room temperature a human body produces ~ 84 W, which increases to ~162 W at a core temperature of 41.8 °C.

Heat losses from the body are often termed sensible (convective, conductive, radiative) and insensible (evaporative, latent). The primary mode of heat loss from the body is by radiation, as described by the Stefan–Boltzmann law,

$$\frac{Q'_{\text{rad}}}{\Delta V} = e_{\text{skin}}\sigma A_{\text{skin}}(T_{\text{skin}} - T_s)^4 \tag{5}$$

where $Q'_{\text{rad}}/\Delta V$ is the power radiated, e_{skin} is the emissivity of the skin (radiating material), σ is Stefan's constant $= 5.6703 \times 10^{-8}$ W·m^{-2}/K, A_{skin} is the skin surface area, T_{skin} is the temperature of the skin (radiator), and T_{s} is the temperature of the surroundings (e.g., air, water, wax). Human skin is a near perfect radiator in the IR, with an emissivity of 0.97. At room temperature, >50% of the heat generated by metabolism is lost by radiation; a clothed adult loses some 50 W at room temperature. This increases to ~66% at a core temperature of 41.8 °C, as is targeted in a number of systemic hyperthermia protocols, when the skin temperature rises to 39–40 °C (3).

Direct transfer of body heat to the molecules around the body (typically air) occurs by conduction, or molecular agitation within a material without any motion of the material as a whole, which is described by Fourier's law,

$$\frac{\Delta Q'_{\text{cond}}}{\Delta V} = \kappa A_{\text{skin}}\frac{\Delta T}{\Delta x} \tag{6}$$

where ΔQ_{cond} is the heat energy transferred per unit volume in time Δt, κ is the thermal conductivity (W·mK^{-1}) of the material surrounding the body (air, water), and ΔT is the temperature difference across thickness Δx of the material. Air is a poor thermal conductor, therefore heat loss by conduction is relatively low. On the other hand, water has a thermal conductivity 20 times that of air at 0 °C, increasing further with temperature, therefore during hyperthermia it is important that any water in contact with the skin is not at a lower temperature. The relative thermal conductivity of body tissues is important in determining thermal conduction within the body from external sources of heat. For example, fat is a relative thermal insulator with a thermal conductivity one third of that of most other tissues, therefore fat bodies are slower to heat.

Convective heat transfer involves material movement and occurs principally via blood moving heat to, or from, the skin and other tissues, and air currents (respiratory and environmental) moving warm air to or from the body. Equation 7 is written for the blood,

$$\frac{\Delta Q'_{\text{conv}}}{\Delta V} = \rho_b c_b[w_b(x,y,z,T)\cdot(T - T_b) + U_b(x,y,z,T)\cdot\nabla T] \tag{7}$$

where w_b is the specific capillary blood flow rate, U_b is the specific blood flow through other vessels. In the context of systemic hyperthermia, where a patient is in a closed chamber, environmental air currents can be minimized. Heat loss by respiration, however, can amount to almost 10% of metabolic heat generation.

Another route of heat loss from the body is evaporation of perspiration from the skin. Because of the very large heat of vaporization of water, cooling of the blood in skin capillaries occurs due to evaporation of sweat. Evaporation from exhaled moisture also results in cooling of the surrounding air.

$$\frac{\Delta Q'_{\mathrm{evap}}}{\Delta V} = m_w \frac{L_v}{\Delta t} \tag{8}$$

where m_w is the mass of the water and L_v is the latent heat of vaporization (2.4×10^6 J·kg^{-1} at 34 °C). In hot conditions with maximal rates of evaporation, heat loss through evaporation of sweat can be as much as 1100 W. Heat loss in the lungs is ~10 W.

Combining the heat generation and heat loss terms leads to a general heat transfer equation, an extension of the classic Pennes bioheat transfer equation.

$$\left[\left(\frac{\Delta Q'_{\mathrm{met}}}{\Delta V} + P_{\mathrm{abs}}\right) - \left(\frac{\Delta Q'_{\mathrm{rad}}}{\Delta V} + \frac{\Delta Q'_{\mathrm{cond}}}{\Delta V} + \frac{\Delta Q'_{\mathrm{conv}}}{\Delta V} + \frac{\Delta Q'_{\mathrm{resp}}}{\Delta V} + \frac{\Delta Q'_{\mathrm{evap}}}{\Delta V}\right)\right] = \rho_{\mathrm{tis}} c_{\mathrm{tis}} \Delta T \tag{9}$$

into which the expressions given in Eqs. 2–8 may be substituted. Precise solution of this equation for temperature distribution is complex and requires a number of simplifying assumptions which have generated significant controversy in bioheat transfer circles. Modeling of temperature distributions within a body subjected to hyperthermia is also complex because of the heterogeneity of thermal characteristics between and within tissue, the directionality of power application, and the dynamic nature of thermoregulation by human body. Nonetheless, the factors governing systemic heating of the body can be appreciated.

INDUCTION OF SYSTEMIC HYPERTHERMIA

Apart from the induction of biological fever by pathogens or toxins, all methods of hyperthermia involve transfer of heat into the body from an external energy source. The required net power to raise the temperature of a 70 kg human from 37 to 41.8 °C (2) is 400 W (5.7 mW). While the heat absorption from these sources is highly nonuniform, distribution of thermal energy by the vascular system quickly results in a uniform distribution of temperature. Indeed, systemic hyperthermia is the only way to achieve uniform heating of tissues. Because physiological thermoregulation mechanisms such as vasodilation and perspiration counteract attempts to increase core body temperature, careful attention must be paid to optimizing the physical conditions for heating such that there is efficient deposition of heat energy in the body and, even more importantly, minimization of heat losses. Wrapping the body in reflective blankets, foil, or plastic film to reduce radiative and evaporative losses, or keeping the surrounding air moist to minimize losses by perspiration are key techniques for achieving a sustained increase in body temperature.

Noninvasive methods of heating include immersing the body in hot water or wax, wrapping the body in a blanket or suit through which heated water is pumped, placing the patient on a heated water mattress, surrounding the body with hot air, irradiating with IR energy, and applying RF or microwave electromagnetic energy. These techniques may be applied singly or in combination. For example, the Pomp–Siemens cabinet used until recently throughout Europe, as well as in the United States, a modification of a device originally developed by Siemens in the 1930s, has the patient lying on a heated water mattress under which an inductive loop generates an RF field, all inside a chamber through which hot air is circulated. The Russian Yakhta-5 system applies a high frequency (13.56 MHz) electromagnetic field through a water-filled mattress to permit whole body heating up to 43.5 °C and simultaneous deep local hyperthermia through additional applicators providing 40.6 MHz electromagnetic radiation. The majority of whole-body hyperthermia systems currently in clinical use employ IR radiation to achieve systemic heating. Invasive approaches to systemic hyperthermia are extracorporeal heating of blood, removed from the body via an arteriovenous shunt, prior to returning it to the circulation, as well as peritoneal irrigation with heated fluid (4). A useful schematic summary of whole-body hyperthermia induction techniques along with references is provided by van der Zee (5).

All of these approaches involve a period of steady temperature increase, followed by a plateau or equilibrium phase where the target temperature is maintained for anywhere from 30 min to several hours, and finally a cool-down phase. Depending on the method of hyperthermia induction, the patient may be anesthetized, consciously sedated, administered analgesia, or not given any kind of medication at all. An epidural block is sometimes given to induce or increase vasodilation. During radiant heat induction, the temperature of the skin and superficial tissues (including tumors) is higher than the core (rectal) temperature whereas during the plateau (maintenance) phase, the skin–superficial tissue temperature drops below the core temperature. As already described, heat losses due to physiological mechanisms limit the rate of heating that can be achieved. When insulation of the patient with plastic foil was added to hot air heating, the heating time to 41.8 °C was decreased from 230 to 150 min (65%), and further to 110 min (48%) by addition of a warm water perfused mattress (5). The homogeneity of the temperature distribution was also significantly increased by the addition of insulation and the water mattress. Noninvasive systemic hyperthermia methodologies typically produce heating rates of 1–10 °C·h^{-1} with 2–3 °C·h^{-1} being most common. More rapid heating can be achieved by the invasive techniques, at the expense of greater risk of infection and morbidity.

COMMERCIALLY AVAILABLE WHOLE-BODY HYPERTHERMIA SYSTEMS

A number of commercially available devices have resulted from the development of these initially experimental systems. The Siemens–Pomp system has already been mentioned, but is no longer commercially available. Similarly, neither the radiant heat chamber developed by Robins (3), and marketed as the Aquatherm system, nor the similar Enthermics Medical Systems RHS-7500 radiant heat device, both producing far IR radiation (IR C) in a moist air chamber, are currently being sold, though they are still in use in several centers. A close relative is the Iratherm2000 radiant heat chamber originally developed by von Ardenne and co-workers (6). In this device, water-filtered infrared radiators at 2400 °C emit their energy from above and below the patient bed, producing near-IR (IR A) radiation that penetrates deeper into tissue than far IR radiation, causing direct heating of the subcutaneous capillary bed. Thermal isolation is ensured by reflective foils placed around the patient. However, note that significant evaporative heat loss through perspiration can be a problem with this system. Also with a significant market share is the Heckel HT 2000 radiant heat device in which patients lie on a bed enclosed within a soft-sided rectangular tent whose inner walls are coated with reflective aluminum foil that ensures that the short wavelength infrared A and B radiation emitted by four radiators within the chamber uniformly bathes the body surface. Once the target temperature is reached, the chamber walls are collapsed to wrap around the body, thereby preventing radiative and evaporative heat loss, and permitting maintenance of the elevated temperature, as shown in Fig. 2.

Another radiant heat device, used mainly in Germany, is the HOT-OncoTherm WBH-2000 whole-body hyperthermia unit which is a chamber that encloses all but the patient's head. Special light-emitting diode (LED) radiators deliver computer-generated, alloy-filtered IR A wavelengths that penetrate the skin to deliver heat to the capillary bed. The manufacturer claims that these wavelengths also preferentially stimulate the immune system. Recently, Energy Technology, Inc. of China has released the ET-SPACE whole-body hyperthermia system, which produces IR A radiation in a small patient chamber into which warm liquid is infused to help increase the air humidity and thereby reduce perspiration losses. A number of low cost, far infrared, or dry, saunas are being sold to private clinics, health clubs, and even individuals for treatment of arthritis, fibromyalgia, detoxification, and weight loss. Examples are the Smarty Hyperthermic Chamber, the TheraSauna, the Physiotherm, and the Biotherm Sauna Dome. Table 2 summarizes features of these commercially available whole-body hyperthermia devices.

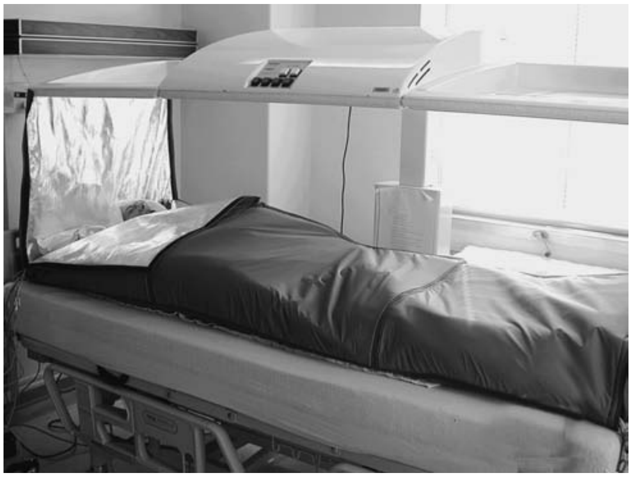


Figure 2. Heckel HT-2000 radiant heat whole body hyperthermia system. Unit at the University of Texas Medical School at Houston. Patient is in the heat maintenance phase of treatment, wrapped in the thermal blankets which form the sides of the chamber during active heating.

BIOLOGICAL EFFECTS OF SYSTEMIC HYPERTHERMIA

An understanding of the biological effects of systemic hyperthermia is critical to both its successful induction and to its therapeutic efficacy. Systemic responses to body heating, if not counteracted, undermine efforts to raise body temperature, while cellular effects underlie both the rationale for the use of hyperthermia to treat specific diseases, and the toxicities resulting from treatment. Although improved technology has allowed easier and more effective induction of systemic hyperthermia, most of the recent clinical advances are due to better understanding and exploitation of specific biological phenomena.

Physiological Effects of Elevated Body Temperature

The sympathetic nervous system attempts to keep all parts of the body at a constant temperature, tightly controlled by a central temperature 'set point' in the preoptic–anterior hypothalamus and a variety of feedback mechanisms. The thermostat has a circadian rhythm and is occasionally reset, for example, during fever induced by infectious agents and endotoxins, but not in endogenously induced hyperthermia. Occasionally, it breaks down completely as in malignant hyperthermia or some neurological disorders affecting the hypothalamus. Ordinarily, when core body temperature rises, the blood vessels initially dilate, heart rate rises, and blood flow increases in an effort to transport heat to the body surface where it is lost by radiation, conduction, and convection. Heart rate increases on average by 11.7 beats$\cdot$min$^{-1}\cdot$ °C^{-1} and typically remains elevated for several hours after normal body temperature is regained. Systolic blood pressure increases to drive the blood flow, but diastolic pressure decreases due to the decreased resistance of dilated vessels, thus there is an increase in cardiac output. Heart rate and blood pressure must therefore be monitored during systemic hyperthermia, and whole-body hyperthermia is contraindicated in most patients with cardiac conditions. Interestingly, hyperthermia increases cardiac tolerance to ischemia/reperfusion injury probably due to activation of manganese superoxide dismutase (Mn-SOD) and involvement of cytokines.

Respiration rate also increases and breathing becomes shallower. Perspiration results in evaporation of sweat from the skin and consequent cooling, while the respiration rate increases in order to increase cooling by evaporation of moisture from expired air. Weight loss occurs despite fluid intake. There is a decrease in urinary output and the urine has a high specific gravity, concentrating urates and

Table 2. **Commercially Available Clinical Whole-Body Hyperthermia Devices**

Manufacturer	Website	Device Name	Heating Mechanism	Temperature Range, °C	Application
Energy Technology	http://www.eti.com.cn/EN/pro/product2.htm	ET-SPACE	Multiple IR radiators (IR A)	39–41.8	Oncology
Heckel Medizintechnik GmbH	http://www.heckel-medizintechnik.de/frameset_e.html	HT 2000 M	4 300W IR radiators (IR A, B)	38.5–40.5	Oncology, rheumatology
Hot-Oncotherm	http://www.hot-oncotherm.com/oncothermia.htm	WBH-2000	Multiple LED radiators (IR A)	37–42	Oncology
Von Ardenne Institut für Angewandte Medizinische Forschung, GmbH	http://www.ardenne.de/med_eng/	Iratherm 800	4 IR radiators (IR A)	37–38	Physical medicine, complementary medicine, oncology
		Iratherm 1000	6 IR radiators (IR A)	37–39	
		Iratherm 2000	10 IR radiators (IR A)	37–42	

phosphates. In endogenously induced hyperthermia, but not in fever, glomerular filtration, as evidenced by the creatinine clearance, decreases with increasing temperature. As already mentioned, metabolic rate increases nonlinearly with temperature, which leads to an increase in blood sugar, decreased serum potassium levels, and increased lactic acid production. All the above normal physiological effects may be enhanced or counteracted by anesthesia or sedation, as well as by disease states such as cancer because of drugs used in treatment or intrinsic pathophysiological consequences of the disease.

At ~42.5 °C, the normal thermocompensatory mechanisms break down and the body displays the symptoms of advanced heat stroke, namely, lack of sweating, rapid heart beat, Cheyne–Stokes breathing, central nervous system disfunction, and loss of consciousness. Ultimately, breathing ceases despite the continuation of a heart beat.

Cellular Thermal Damage

When temperature is increased by a few degrees Celcious, there is increased efficiency of enzyme reactions (Arrhenius equation), leading to increased metabolic rates, but at temperatures $> 40\ ^\circ C$ molecular conformation changes occur that lead to destabilization of macromolecules and multimolecular structures, for example, to the side chains of amino acids in proteins, which in turn inhibit enzyme action. Small heat shock proteins (HSP) interact with the unfolding proteins to stabilize them and prevent their aggregation and precipitation. Eventually, however, at ~42 °C, complete denaturation of proteins begins that totally disrupts many molecular processes, including deoxyribonucleic acid (DNA) repair. Thus systemic hyperthermia can have significant effects when paired with drugs that cause DNA damage (e.g., for chemotherapy of cancer).

Membranes are known to be extremely sensitive to heat stress because of their complex molecular composition of lipids and proteins. At a certain temperature, lipids change from the tightly packed gel phase to the less tightly packed liquid crystalline phase, and permeability of the cell membrane (membrane fluidity) increases. As temperature increases further, the conformation of proteins also becomes affected, eventually resulting in disorderly rearrangement of the lipid bilayer structure and receptor inactivation or loss. Temperature changes of ~5 °C are necessary to cause measurable changes in normal cell membrane permeability. Heat-induced cell membrane permeability can be exploited to increase drug delivery, for example, transdermally, or into tumor cells. Increased vascular permeability due to thermal increase of endothelial gap size also aids drug delivery into tumors. At higher temperatures, heat damage to membranes can cause cell death, but it will also interfere with therapeutic approaches that depend on membrane integrity (e.g., receptor targeted drug delivery, antibodies, etc.). Irreversible disruption of cytoplasmic microtubule organization and eventual disaggregation, as well as disruption of actin stress fibers and vimentin filaments, occur at high temperatures (43–45 °C) above those used in whole-body hyperthermia, but these cytoskeletal effects are of concern with loco-regional hyperthermia.

A variety of effects in the cell nucleus also occur at high temperatures (>41 °C) including damage to the nuclear membrane, increases in nuclear protein content, changes in the structure of nucleoli, inhibition of DNA synthesis and chromosomal damage in S-phase. These changes in nuclear structure compromise nuclear function and may cause cell death, though they are unlikely to be significant at the temperatures achieved in systemic hyperthermia. Disaggregation of the spindle apparatus of mitotic cells may be responsible for the high thermal sensitivity of cells in mitosis, as well as in S phase. Hyperthermic inactivation of polymerase β, an enzyme primarily involved in DNA repair, is sensitized by anesthetics and may have a role to play in the enhancement of the effects of ionizing radiation by systemic hyperthermia, as well as in augmenting the cytotoxic effect of drugs that cause DNA damage.

Metabolic Effects

Moderate increases in temperature lead to increased cellular reaction rates, which may be seen as increased oxygen consumption and glucose turnover. In consequence, cells may become deprived of nutrients, the intracellular ATP concentration falls, accumulation of acid metabolites increases pH, and thermal sensitivity increases. Such conditions are found in tumors and may contribute to their sensitivity to heat. Further acidifying tumor cells during hyperthermic treatment seems a promising approach as is discussed further below. At high temperatures, the citric acid cycle may be damaged leading to other acidic metabolites. Increased plasma acetate has been measured following clinical whole-body hyperthermia treatments, which reduces both release of fatty acids from adipose tissue into plasma and subsequent lipid oxidation.

Endocrine Function

Increases in plasma levels of an array of hormones have been noted after whole-body hyperthermia. Increased ACTH levels appear to be accompanied by increased levels of circulating endorphins. This may explain the sense of well-being felt by many patients after systemic hyperthermia treatment, and the palliative effect of hyperthermia treatments for cancer. Increased secretion of somatotropic hormone after systemic hyperthermia has also been measured (7).

Thermal Tolerance

Thermal tolerance is a temporary state of thermal resistance, common to virtually all mammalian cells, which develops after a prolonged exposure to moderate temperatures (40–42 °C), or a brief heat shock followed by incubation at 37 °C, and also certain chemicals. The decay of thermotolerance occurs exponentially and depends on the treatment time, the temperature, and the proliferative status of the cells. Several days are usually required for baseline levels of heat sensitivity to be regained, which has important implications for fractionated therapy. When pH is lowered, less thermal tolerance develops, and its decay is slower. Thus the long periods at moderate temperature achieved by clinical systemic hyperthermia systems should

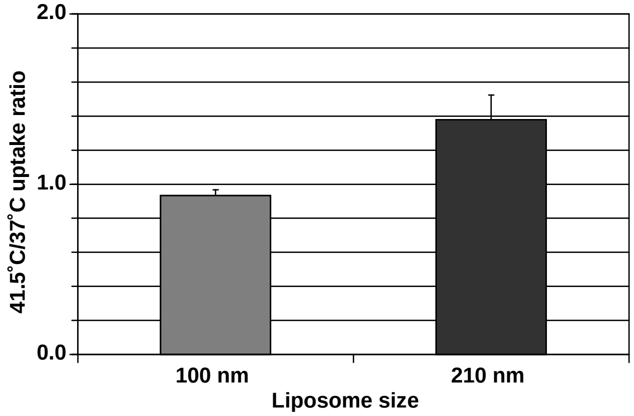


Figure 3. Increase in tumor uptake of large liposomes after 1 h of 41.5 °C whole-body hyperthermia. Systemic heat treatment increased the effective pore size from ~210 to 240 nm. Because of the large pore size in MTLn3 tumors, 100 nm (average diameter) liposomes were able to pass into the tumor equally well at normal and elevated temperatures. The increased effective pore size due to hyperthermia allowed larger 200 nm liposomes, which were partially blocked at normal temperatures, to pass more effectively into the tumor.

induce thermal resistance in normal cells, while the acidic parts of tumors should be relatively unaffected. This has not, however, been studied clinically. The mechanisms involved in the induction of thermotolerance are not well understood, but there is mounting evidence that heat shock proteins are involved.

Step-Down Sensitization

Another distinct phenomenon is step-down sensitization in which an exposure of cells to temperatures >43 °C results in increased sensitivity to subsequent temperatures of 42 °C or lower. This can be important clinically for local and regional hyperthermia if there are marked variations in temperature during the course of treatment, the magnitude of the effect depending on the magnitude of the temperature change. It has been suggested that this phenomenon could be exploited clinically by administering a short, high temperature treatment prior to a prolonged treatment at a lower temperature, thereby reducing pain and discomfort. Since temperatures >43 °C cannot be tolerated systemically, a local heat boost would be required to take advantage of this effect for whole body hyperthermia. So far, there is no evidence that tumor cells are differently sensitized by step-down heating than normal cells.

Effect of Hyperthermia on Tumors

It was initially thought that tumor cells have intrinsically higher heat sensitivity than normal cells, but this is not universally true. Although some neoplastic cells are more sensitive to heat than their normal counterparts, this appears to be the case at temperatures higher than those used in systemic hyperthermia. Tumors *in vivo*, on the other hand, often do have a higher thermal sensitivity than normal tissues because of abnormal vasculature (reduced blood flow), anaerobic metabolism (acidosis), and nutrient depletion. Due to their tortuous and poorly constructed vasculature, tumors have poor perfusion, thus heat dissipation by convection is reduced. At high temperatures (43 °C and up) this means that tumors become a heat reservoir with a consequent rise in temperature, which if maintained for too long damages the microcirculation and further impairs convective heat loss. Also increased fibrinogen deposition at damaged sites in the vascular wall leads to clusion of tumor microvessels. Significant heating of the tumor cells results, which may be directly cytotoxic. Additionally, the impaired blood flow brings about acidosis, increased hypoxia and energy depletion all of which increase the heat sensitivity of tumor cells (8). At lower temperatures, typical of those achieved in whole-body hyperthermia, blood flow increases (9) though the mechanism is not well understood. For these reasons, along with the historical evidence for antitumor effects of fever and the metastatic nature of malignant disease, cancer has become the main focus of systemic hyperthermia.

Systemic hyperthermia results in increased delivery of drugs to tumor sites because of increased systemic blood flow. It can also increase blood vessel permeability by increasing the effective pore size between the loosely bound endothelial cells forming tumor microvessels, permitting larger molecules, such as nanoparticles and gene therapy vectors, to pass into the interstitium (10). Figure 3 shows increased uptake of 210 nm liposomes in rat breast tumors after 1 h of 41.5 °C whole-body hyperthermia. Heat may also be toxic to endothelial cells, resulting in a transient normalization of vascular architecture and improvement in blood flow (11). Another barrier to drug delivery is the high interstitial pressure of many tumors. Since whole-body hyperthermia, even at fever-range temperatures, causes cell death (apoptosis and necrosis) within tumors it reduces the oncotic pressure allowing greater penetration of large molecules. Table 3 summarizes the interactions of systemic hyperthermia which facilitate nanoparticle delivery to tumors.

Table 3. Whole-Body Hyperthermia Facilitates Nanoparticle Therapy

Heat Interaction	Therapeutic Effect
↑ Blood flow	↑ Nanoparticle delivery to tumor
↑ In endothelial gap size	↑ Nanoparticles in interstitium
↑ Endothelial cell apoptosis/necrosis → transient normalization of vasculature	↑ Nanoparticles in interstitium
↑ Tumor cell apoptosis/necrosis ↓ oncotic pressure	↑ Nanoparticles in interstitium
Temperature-dependent ↑ in permeability of liposome bilayer	↑ And synchronization of drug release
Cellular and molecular effects in tumor	↑ Drug in tumor cell ↑ drug efficacy
Direct interactions with drug	↑ Drug efficacy

Whole-Body Hyperthermia and the Immune System

An increase in ambient temperature can serve as a natural trigger to the immune system and it appears that the thermal microenvironment plays a critical role in regulating events in the immune response. The early work of Coley on cancer therapy with infectious pyrogens implicated fever-induced immune stimulation as the mediator of tumor responses (1). While there have been numerous *in vitro* studies of the effect of temperature on components of the immune system, indicating that the thermal milieu regulates T lymphocytes, natural killer (NK) cells, and dendritic cells (DC), *in vivo* examinations of the immune effects of systemic hyperthermia are relatively few. Initial animal model studies concluded that whole-body hyperthermia resulted in immunosuppression, but high temperatures were used, tumors were mostly immunogenic, and immune response was merely inferred from the incidence of metastatic spread rather than from measurement of specific markers of immune system activation. The majority of *in vivo* studies in animals provide evidence of a nonspecific host reaction in response to hyperthermia in which both T and B lymphocytes, as well as macrophages, are involved (12). Although NK cells are intrinsically more sensitive *in vitro* to heat than B and T cells, their activation by systemic hyperthermia has been observed. Microwave induced whole-body hyperthermia of unrestrained, unanesthetized mice at 39.5–40 °C for 30 min, three or six times weekly, resulted in increased NK cell activity and reduced pulmonary metastasis in tumor-bearing mice, but none in normal mice (13). Evidence for hyperthermia-induced human tumor lysis by IL-2 stimulated NK cells activated by HSP72 expression also exists (14). Increased numbers of lymphocyte-like cells, macrophages, and granulocytes are observed in the tumor vasculature and in the tumor stroma of xenografts and syngeneic tumors in mice immediately following a mild hyperthermia exposure for 6–8 h. In the SCID mouse/human tumor system tumor cell apoptosis seen following treatment was due largely to the activity of NK cells. The investigators hypothesize heat dilatation of blood vessels and increased vessel permeability may also give immune effector cells greater access to the interior of tumors (15). In balb/C mice, fever-range whole-body hyperthermia increased lymphocyte trafficking, resulting in early responsiveness to antigen challenge (16). Thus systemic hyperthermia may be an effective, nontoxic adjuvant to immunotherapy.

A recent clinical study examined the effect of whole-body hyperthermia combined with chemotherapy on the expression up to 48 h later of a broad range of activation markers on peripheral blood lymphocytes, as well as serum cytokines and intracellular cytokine levels in T cells, and the capacity of these cells to proliferate. Immediately after treatment with 60 min of 41.8 °C WBH as an adjunct to chemotherapy, a drastic but transient, increase in peripheral NK cells and CD56+ cytotoxic T lymphocytes was observed in the patients' peripheral blood. The number of T cells then briefly dropped below baseline levels, a phenomeonon that has also been observed by others (17). A marked, but short-lived, increase in the patients' serum levels of interleukin-6 (IL-6) was also noted. Significantly increased serum levels of tumor necrosis factor-alpha (TNF-alpha) were found at 0, 3, 5 and 24 h posttreatment. Further immunological consequences of the treatment consisted of an increase in the percentage of peripheral cytotoxic T lymphocytes expressing CD56, reaching a maximum at 48 h post-WBH. Furthermore, the percentage of CD4+ T cells expressing the T cell activation marker CD69 increased nearly twofold over time, reaching its maximum at 48 h. Since similar changes were not observed in patients receiving chemotherapy alone, this study provided strong evidence for prolonged activation of human T cells induced by whole-body hyperthermia combined with chemotherapy (18).

Activation of monocytes has been observed following hot water bath immersion such that response to endotoxin stimulation is enhanced with concomitant release of TNF-α. Macrophage activation and subsequent lysosomal exocytosis were observed in the case of a patient treated for liver metastases by hyperthermia. Lysosomal exocytosis induced by heat may be an important basic reaction of the body against bacteria, viruses, and tumor growth and was proposed as a new mechanism of thermally induced tumor cell death mediated by an immune reaction (19).

Several investigators have suggested that the immune changes seen during *in vivo* whole-body hyperthermia are mediated by elevations in the plasma concentrations of either catecholamines, growth hormone, or beta-endorphins. In volunteers immersed in a heated water bath, neither recruitment of NK cells to the blood, nor the percentages or concentrations of any other subpopulations of blood mononuclear cells were altered by hormone blockade. However, somatostatin partly abolished the hyperthermia induced increase in neutrophil number. Based on these data and previous results showing that growth hormone infusion increases the concentration of neutrophils in the blood, it was suggested that growth hormone is at least partly responsible for hyperthermia induced neutrophil increase. A similar study suggested that hyperthermic induction of T lymphocytes and NK cells is due to increased secretion of somatotropic hormone (7).

The peripheral blood level of prostaglandin E_2 (PGE_2), which may act as an angiogenic switch, transforming a localized tumor into an invasive one by stimulating new blood vessel growth, and which also has an immunosuppressive effect, is elevated in patients with tumors compared to healthy control subjects. In a clinical study of cancer patients receiving 1–2 h of 41.8–42.5 °C whole-body hyperthermia, or extracorporeal hyperthermia, blood levels of PGE_2 decreased markedly after treatment and correlated with tumor response (20).

In addition to their role as protectors of unfolding proteins, extracellular heat shock proteins (HSP) can act simultaneously as a source of antigen due to their ability to chaperone peptides and as a maturation signal for dendritic cells, thereby inducing dendritic cells to cross-present antigens to CD8+ T cells (21). Heat shock proteins can also act independently from associated peptides, stimulating the innate immune system by eliciting potent proinflammatory responses in innate immune cells. The heat shock response also inhibits cyclooxygenase-2 gene expression at the transcriptional level by preventing the activation of nuclear factor-kappaB (NFκB) (22). Thermal upregulation of HSPs (HSP70 and HSP110) is strongest in lymphoid tissues and may relate to the enhanced immune responses that are observed during febrile temperatures. It has been proposed that local necrosis induced by hyperthermic treatment induces the release of HSPs, followed by uptake, processing and presentation of associated peptides by dendritic cells. By acting as chaperones and as a signal for dendritic cell maturation, HSP70 might efficiently prime circulating T cells. Therefore, upregulating HSP70 and causing local necrosis in tumor tissue by hyperthermia offers great potential as a new approach to directly activate the immune system, as well as to enhance other immunotherapies (23,24).

CLINICAL TOXICITIES OF WHOLE-BODY HYPERTHERMIA TREATMENT

At fever-range temperatures, adverse effects of systemic hyperthermia treatment are minimal however, at higher temperatures they can be significant, even fatal. On the other hand, the teratogenic effects (birth defects, still births, spontaneous abortions) and °Cular damage (cataract induction) resulting from electromagnetic fields used in local hyperthermia are not seen in systemic hyperthermia. The transient cardiorespiratory effects of elevated temperature can, however, lead to severe toxicity. Elevated heart rate, especially at high temperatures may result in arrythmias or ischemic heart failure, consequently patients have to be very carefully screened with regard to their cardiac status. Beta blockade has generally been found to be deleterious although infusion of esmolol has been safely carried out (25). Pulmonary hypertension and edema due to capillary leak may also be seen, but like the cardiac effects, these return to baseline a few hours after treatment. Increased serum hepatic enzymes have been noted, but these may be cancer related. All these toxicities are less prevalent or less severe with radiant heat systems, particularly at lower temperatures, and when light conscious sedation is used rather than general anesthesia. For example, decreased platelet count, decreased plasma fibrinogen, and other factors leading to increased blood clotting have been noted, particularly in extra-corporeal hyperthermia, but also with other methods of heating carried out under inhalation-administered anesthesia drugs. On the other hand, with whole-body hyperthermia under conscious sedation there is no evidence of platelet drops (26) and animal studies even show platelet stimulation providing protection against radiation induced thrombocytopenia.

Since systemic hyperthermia is almost never used as a single treatment modality, it is important to recognize that whole-body hyperthermia combined with radiation and chemotherapy can enhance some of the toxicities associated with these modalities. For example, the cardiotoxicity of doxorubicin and both the renal toxicity and hematological toxicity of platinum agents may increase under hyperthermia (27), while the muscle and peripheral nervous system effects of radiation and some drugs can also be enhanced (28). Bone marrow suppression is the limiting toxicity of many chemotherapy drugs but there is little data to suggest that whole body hyperthermia exacerbates this effect. On the contrary, the synergy of hyperthermia with several chemotherapy agents may mean that lower doses can be used, resulting in less toxicity. For example, systemic hyperthermia combined with carboplatin achieves therapeutic results without elevation of myelosuppression and responses have occurred at lower than normal doses (29). Pressure sores can easily develop at elevated temperatures thus care must be taken not only in patient placement and support, but also with application of monitoring devices. If heat dissipation is locally impaired, for example, at pressure points, hot spots occur that can lead to burns. This is rarely a problem with fever-range whole-body hyperthermia, but in anesthetized patients undergoing high heat regimens burns are not uncommon.

Following systemic hyperthermia treatments, malaise and lethargy are almost universally experienced although these may be counteracted by pain relief and a sense of well-being due to released endorphins. However, the faster the target temperature is reached, the less the exhaustion (6), thus attention to minimizing heat dissipation during the heat-up phase and using efficient heating devices, such as those that generate heat by several mechanisms (e.g., radiant heat and EM fields), add a regional heat boost, or produce near-IR radiation that is preferentially absorbed, is advantageous to patient well being. Fever after treatment in the absence of infectious disease is not uncommon and may be associated with an inflammatory response to tumor regression. Nausea and vomiting during the first couple of days after treatment are also common. Outbreaks of herpes simplex (cold sores) in susceptible individuals have also been noted, but are easily resolved with acyclovir.

THERMAL DOSE

The definition of dose for systemic hyperthermia is problematic. An applied dose would be the amount of heat energy generated or delivered to the body but even if it can be measured, this quantity does not predict biological effects. By analogy with ionizing radiation, the absorbed dose would be amount of thermal energy absorbed per unit mass of tissue ($J \cdot kg^{-1}$), however, this is not a quantity that can be readily measured, or controlled, neither would it necessarily predict biological effects. As indicated in the previous sections, the effects of systemic hyperthermia depend on (*1*) the temperature, and (*2*) the duration of heating, but not on the energy required to produce the temperature rise. This leads to the concept of time at a given temperature as a practical measure of dose. In reality, however, temperature is seldom constant throughout a treatment, even in the plateau phase of systemic hyperthermia, so time at temperature is at best a crude measure. Nonetheless, it is the one that is used most often clinically for whole-body hyperthermia because of its simplicity. Ideally, the dose parameter should allow for comparison of treatments at different temperatures. Based on the Arrhenius relationship and measured cell growth inhibition curves, the heating time at a given temperature relative to the heating time at a standard temperature or thermal dose equivalent (TDE), was defined empirically as,

$$T_1 = t_2 \cdot R^{(T_1 - T_2)} \tag{10}$$

A discontinuity occurs in the temperature-time curves between 42 and 43 °C for both cells in culture and heated tissues, thus the value of R changes for temperatures above the transition: $R \sim 2 < 42.5$ °C and $R \sim 5 > 42.5$ °C *in vitro* while for *in vivo* heating studies, $R = 2.1$ below the transition temperature and 6.4 above 42.5 °C. In practice, a finite time is required for the body or tissue of interest to reach the target temperature, temperature fluctuates even after the target temperature is reached, and there is a cooling period after heating ceases. If the temperature is measured frequently throughout treatment, the temperature–time curves can be integrated to provide the accumulated thermal dose that produces an equivalent effect to that resulting from holding the cells–tissue at a constant reference temperature for a given a period of time:

$$t_{43} = \int_{t_i}^{t_f} R^{43-T(t)} dt \tag{11}$$

where t_{i} and t_{f} are the initial and final times of the heating procedure (30). This thermal isoeffect dose (TID) is usually expressed in minutes is sometimes known as the tdm43 or the cumulative equivalent minutes (CEM 43 °C). While a biological factor has now been built in to the dose measure, and the integrated TID allows for temperature variations during heat-up and cool-down phases, it does not take into account thermal tolerance and step-down sensitization. Nor is it particularly relevant to clinical whole-body hyperthermia where multiple physical and biological effects combine in a complex manner although for a given patient, time–temperature profiles are generally reproducible from one treatment to another. A further modification attempts to take into account temperature inhomogeneity through the measurement of temperature at multiple sites and defining T90, namely, that temperature exceeded by 90% of the measurements (or correspondingly 20%: T20; or 50%: T50). The TID is then expressed as cumulative equivalent minutes that T90 is equal to 43 °C (CEM 43 °C T90) (31).

The efficiency of adjuvant hyperthermia in enhancing the biological effectiveness of other treatments is often reported in terms of the thermal enhancement factor (TEF) or thermal enhancement ratio (TER). This quantity is defined in terms of the isoeffect dose as,

$$\mathrm{TER} = \frac{\text{dose of treatment to achieve a given endpoint}}{\text{dose of treatment with heat to achieve the same endpoint}} \tag{12}$$

In clinical and laboratory studies, the TER is often computed on the basis of isodose rather than isoeffect, for example, in the case of hyperthermia plus drug induced arrest of tumor growth, $\mathrm{TER} = \mathrm{TGD_{HT}/TGT_{RT}}$, where $\mathrm{TGD_{HT}}$ is the tumor growth delay due to hyperthermia plus chemotherapy, and $\mathrm{TGT_{RT}}$ is the tumor growth delay resulting from chemotherapy at room temperature. Similarly, the enhancing effect of hyperthermia on radiation treatment may be expressed through $\mathrm{TER} = \mathrm{D0_{HT}/D0_{RT}}$ or $\mathrm{TER} = \mathrm{LD50_{HT}/LD50_{RT}}$, where D0 is the time required to reduce survival to $1/e$ of its initial value, and LD50 is the lethal dose to 50% of cells.

TEMPERATURE MEASUREMENT

Since systemic hyperthermia achieves a uniform temperature distribution, except for possible partial sanctuary sites, thermometry for systemic hyperthermia is much less challenging than for regional or intracavitary hyperthermia, but it is still important to prevent adverse effects, especially burns. Also, convection can induce steep thermal gradients, especially around major blood vessels, so that careful placement of temperature probes is required. Most practitioners of whole-body hyperthermia measure temperature in several locations, typically the rectum, the esophagus, and at several skin sites. During heat-up, the esophageal temperature is usually 1–2 °C higher than the rectal temperature, but during plateau phase it drops to 0.5–1.5 °C below the rectal temperature. Continuous and accurate temperature measurement is particularly important when temperatures >41°C are to be achieved, as critical, life-threatening changes can occur in minutes or even seconds and over changes in temperature of as little as 0.1–0.2 °C because of the nonlinear response to temperature. For moderate temperature systemic hyperthermia, temperature measurement to within 0.1 °C is usually adequate, but a precision of 0.01 °C is desirable when heating to >41 °C and also allows determination of the specific absorption rate from the slope of the temperature versus time curve. The temperature measuring device must be insensitive to all other influences, such as ambient temperature, moisture, nearby electromagnetic fields, and so on and satisfying this criterion can be difficult. Frequent calibration of thermometers in the working range of temperatures is important since some thermometers appear fine at 30 °C, but drift substantially at 40 °C and above. Stringent quality control of any thermometry system is required to monitor accuracy, precision, stability, and response time.

Table 4 summarizes the different types of thermometer probes available for internal and external body temperature measurements, and their relative merits and disadvantages for systemic hyperthermia. Thermistors are most often used for standard temperature monitoring sites while thermocouples are used for tumor or other intra-tissue measurements. Recently, noninvasive methods of temperature measurement have been developed that are beginning to see application in hyperthermia. Thermography provides a two-dimensional (2D) map of surface temperature by measurement of infrared emission from the body, though deep-seated hot structures may be visualized because of heat carried by blood flow from the interior heat source to the skin. It is useful to detect skin hotspots and therefore in burn prevention. Since temperature-induced changes in the mechanical properties of tissue lead to altered ultrasound propagation velocity, mapping of ultrasound velocity can also provide a visual map of temperature. Tomographic reconstruction of 2D or 3D temperature is theoretically possible, but it is difficult in practice because of the heterogeneity of tissue characteristics. A

Table 4. Temperatures Probes for Systemic Hyperthermia

Probe Type	Measurement Principle	Accuracy	Sensitivity	Stability	Advantages or Disadvantages
Clinical	Expansion of mercury or alcohol in glass	Moderate ≤ 0.1 °C	Low	High	Large size, inflexible. Slow response.
Platinum resistance thermometer	Linear resistance change with temperature	High ~ 0.02 °C		Used as standard for calibration of other types of thermometers.	Expensive. Difficult to calibrate. Large size. Sensitive to shock.
Thermocouple	Seebeck effect: temperature dependent voltage difference between two conductors made of different metals	Moderate ≤ 0.1 °C	Moderate to high	Moderate	*Small sensor*. Nonlinear voltage change with temp. Sensitive to EM fields. Can't handle steep temp. gradients.
Thermistor (e.g., Bowman Loop Larsen probe)	Inverse relationship between temperature and semiconductor resistance	High < 0.05 °C	High	Poor Require frequent recalibration	*Short time constant*. Not interchangeable. Sensitive to EM fields.
GaAs	Temperature specific absorption	Moderate	Low		*Small size*
Optical (fiber optic probe):	Change w/temp.:				*Not sensitive to EM fields*. *Small size*.
LCD birefringent crystal fluorescent phosphor	Color reflectance		Low	Low	Unstable
	Refraction of polarized light		Low		
	Decay of fluorescence	Very high			

Table 5. Summary of Clinical Trials of Whole-Body Hyperthermia[a]

First Author	Public Year	Study Type	Number of Patients	Disease	Protocol	Result of WBH	Reference, PMID
WBH Alone							
Kraybill, W.G.	2002	Phase I		Advanced solid tumors	3–6 h at 39.5–40.0 °C	Well tolerated No significant adverse events ↓ in circulating lymphocytes	16, 12028640
Steinhausen, D.	1994	Phase I	103	Advanced refractory or recurrent cancers	1 h at 41.8 °C + hyperglycemia + hyperoxemia	Minimal side effects 52 responses (50%)	8023241
WBH + Chemotherapy							
Bakshandeh, A.	2003	Phase II	25	Nonmetastatic malignant pleural mesothelioma	1 h at 41.8 °C + ifosfamide + carboplatin + etoposide	Grade III/IV neutropenia and thrombocytopenia 5 partial remissions (20%)	12609573
Bull, J.M.	1992	Phase II	17	Advanced metastatic sarcoma	2 h at 41.8–42.0 °C + BCNU	Limiting toxicity = thrombocytopenia 7 responses/SD (41%) ↑ survival	33, 1607734
Bull, J.M.	2002	Phase I	13	Various chemotherapy resistant cancers	6 h at 40.0 °C + doxil + 5-FU + metronomic interferon-α	Grade III toxicities 9 responses/SD (69%)	60
Bull, J.M.	2004	Phase I	33	Advanced metastatic cancers (GI, breast, head and neck, sarcoma, neuroendocrine)	6 h at 40.0 °C + cisplatin + gemcitabine + interferon-α	20 responses/SD (66%) ↑ survival ↑ quality of life	35
Douwes, F.	2004	Pilot	21	Ovarian cancer	1-2 h at 41.5–42.0 °C + cisplatin or carboplatin + hyperglycemia	18 responses/SD (86%) ↑ quality of life	15108039
Engelhardt, R.	1990	Pilot	23	Advance metastatic melanoma	1 h at 41.0 °C + cisplatin + doxorubicin	Slight ↑ in myelotoxicity 10 responses/SD Response rate = that in literature for chemo alone	52, 2198312
Guan, J.	2005	Phase II	32	Advanced cancers		94% responses/SD Pain reduction in all pts. Increased KPS Decreased tumor markers	65
Hegewisch-Becker, S.	2002	Phase II	41	Pretreated advanced metastatic colorectal cancer	1 h at 41.8 + oxaliplatin + leucovorin + 5FU	No excess toxicity 31 responses/SD (76%)	55, 12181242
Hildebrandt, B.	2004	Phase I/II	28	Metastatic colorectal cancer	1 h at 41.8–42.1 °C + hyperglycemia + hyperoxemia + folinic acid + 5-FU + mitomycin C	Grade III/IV toxicities 11 responses/SD (39%)	44, 15204528

Hou, K.	2004	Phase II	54	Advanced cancers	1–2 h at 41.8–42.5 °C, extracorporeal + chemotherapy vs. chemotherapy alone	75.3% responses/SD 72.6% ↓ tumor markers 70% pain relief improved sleep ↑ weight, appetite, KPS All signifly > control Reversible toxicities	68
Ismael-Zade, R.S.	2005	Pilot	5	Pediatric renal cell carcinoma	3 h at 41.8–42.5 °C + doxorubicin + interferon-α	No complications 5 responses (100%)	15700247
Kurpeshev, O.K.	2005	Phase II	42	Various disseminated cancers	1–2 h at 41.0–42.3 °C + poly-chemotherapy	Regression of metastases. Pain reduction	66
Richel, O.	2004	Phase II	21	Metastatic and recurrent cervical cancer	1 h at 41.8 °C + carboplatin	Grade III/IV leucopenia, thrombopenia, anemia, renal toxicity 16 responses/SD (76%)	57, 15581981
Robins, H.I.	1993	Phase I	30	Various refractory cancers	1 h at 41.8 °C + carboplatin	Myelotoxicity 9 responses (30%)	53, 8355046
Robins, H.I.	1997	Phase I	16	Various refractory cancers	1 h at 41.8 °C + L-PAM	Lower platelet nadir myelosuppression 8 responses/SD (50%)	48, 8996137
Strobl, B.	2004	Phase II	7	Metastatic cervical cancer	1 h at 41.5–41.8 °C + paclitaxel + carboplatin	Grade II alopecia Grade III/IV thrombopenia, neutropenia ↑ survival	58
Westermann A.M.	2001	Phase II	14	Platinum resistant ovarian cancer	1 h at 41.8 °C + carboplatin	Grade IV thrombocytopenia, grade III neutropenia 9 responses/SD (64%)	11378341
Westermann A.M.	2003	Phase II	95	Metastatic sarcoma	1 h at 41.8 °C + ifosfamide + carboplatin + etoposide	Neutropenia, thrombocytopenia, infection 58 responses/SD (61%)	50, 12759526
WBH + Radiation							
Overgaard, J.	1995	Randomized multicenter	70	Metastatic melanoma	1 h at 43 °C + fractionated RT vs. FRT alone	Improved local tumor control, ↑ survival	41, 7776772
Robins, H.I.	1990	Pilot	8	Nodular lymphoma, chronic lymphocytic leukemia	41.8 °C + TBI vs. LON + TBI	8 responses/SD ↑ survival	24, 2182581

[a] Published since 1990.

number of magnetic resonance (MR) techniques have been used for thermal mapping and BSD Medical and SIEMENS Medical Systems have collaborated to develop a hybrid hyperthermia/MRI system, although it is not a whole-body hyperthermia machine. Currently, the most widely accepted MR technique is the proton resonance frequency (PRF) method that exploits the temperature dependence of the chemical shift of water. Unlike the value of the water spin-lattice relaxation time or the molecular diffusion coefficient, both of which have been used for MRI temperature measurements, the thermal coefficient relating temperature to the water chemical shift has been shown to be essentially independent of tissue type and physiological changes induced by temperature (32). Recently an interleaved gradient echo–echo planar imaging (iGE-EPI) method for rapid, multiplanar temperature imaging was introduced that provided increased temperature contrast-to-noise and lipid suppression without compromising spatio-temporal resolution (33).

CLINICAL EXPERIENCE

Cancer

Systemic hyperthermia has been used mostly for treatment of cancer because of its potential to treat metastatic disease. Initial treatments aimed to produce direct killing of tumor cells based on the premise, now understood not to be universally true, that cancer cells are more susceptible to elevated temperatures than normal cells, and the higher the temperature the greater the tumor cell kill. Maximally tolerated temperatures of 41.5–42 °C were therefore maintained for 1–2 h as the sole treatment. Response rates were, however, disappointing. Tumor regressions were observed in less than half the cases, no tumor cures were achieved, and remissions were of short duration. It became apparent that the heterogeneity of cell populations within tumors, along with micro-environmental factors, such as blood/nutrient supply, pH, and oxygen tension prevent the thermotoxic results achieved in the laboratory. Consequently, the focus of research on systemic hyperthermia shifted to using hyperthermia as an adjunct to other cancer therapies, principally chemotherapy and radiotherapy. It is important to note that because of the experimental status of systemic hyperthermia treatment for cancer, almost all clinical trials, summarized in Table 5, have been performed on patients with advanced disease for whom whole-body hyperthermia, either as a sole therapy, or as an adjunct, is a treatment of last resort. In these cases, any response whatsoever is often remarkable. Nonetheless, a number of hyperthermia centers in Europe have discontinued systemic hyperthermia because the high temperature protocols required intensive patient care and led to unacceptable toxicities, especially in light of the efficacy and reduced toxicities of newer generation chemotherapies. Large, randomized, multicenter, Phase III trials are, however, needed to firmly establish the benefits of systemic hyperthermia in conjunction with chemotherapy and radiation. Also, validation and optimization of fever-range temperature protocols are much needed.

Systemic Hyperthermia and Chemotherapy. The beneficial interaction of hyperthermia with several classes of chemotherapy agents, acting via several mechanisms as summarized in Table 6, has spurred a variety of thermo-chemotherapy regimens and several clinical trials of systemic hyperthermia and chemotherapy are ongoing. While the results have been mixed, elevated response rates were recorded in the treatment of sarcoma when systemic hyperthermia was combined with doxorubicin and cyclophosphamide (54) or BCNU (34). Systemic hyperthermia is the only way to heat the lung uniformly, and impressive response rates and increased durations of response have been achieved in both small cell and nonsmall cell lung cancer treated with the combination of whole body hyperthermia at 41 °C for 1 h with adriamycin, cyclophosphamide, and vincristine (ACO protocol) (34). Neuroendocrine tumors also appear to have increased sensitivity to systemic hyperthermia and multidrug chemotherapy (51).

Optimal combination of whole-body hyperthermia with chemotherapy requires an understanding of the mechanisms of interaction of heat with individual drugs or drugs in combination. Preclinical data is consistent with the concept that the timing of chemotherapy during whole-body hyperthermia should affect therapeutic index. For example, Fig. 4 shows the effect on tumor cures in mammary carcinoma bearing rats of 6 h of 40 °C whole-body hyperthermia administered with, or 24 or 48 h after gemcitabine. A synergistic response was obtained when hyperthermia was begun with gemcitabine administration or 48 h later. The effect of gemcitabine was completely negated, however, when hyperthermia was administered 24 h after the start of heating, perhaps due to cell cycle effects. With cisplatin, the greatest therapeutic index is achieved if the drug is given 24 h before the start of whole-body hyperthermia, thereby preventing thermal augmentation of cisplatin induced nephrotoxicity (55). In a clinical investigation of multiple cycles of radiant heat whole-body hyperthermia combined with carboplatin, Ifosfamide, etoposide, and granulocyte colony stimulating factor, it was found that toxicity was minimized when carboplatin was

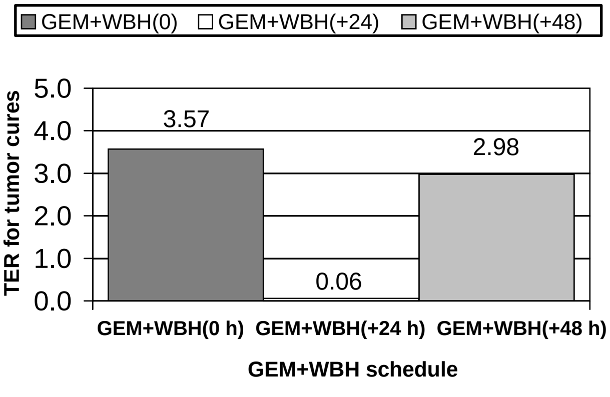


Figure 4. Schedule dependence of fever range whole-body hyperthermia enhanced gemcitabine tumor cures. A supraadditive cure rate occurred when whole-body hyperthermia (WBH) was given at the same time as gemcitabine administration or 48 h later. When hyperthermia followed gemcitabine by 24 h the number of cures dropped to almost zero, well below the number achieved with gemcitabine alone.

Table 6. Chemotherapy Agents Used with Whole-Body Hyperthermia

Class of Agent	Likely Mechanism of Heat Interaction	Drugs Used with WBH in Clinical Studies	Investigator References[a]
Alkylating agents	Impaired DNA repair	Cyclophosphamide (CTX)	Parks, 1983 (4)
	Improved pharmacokinetics		Engelhardt, 1988 (34)
		Dacarbazine (DTIC)	Lange, 1983 (35)
		Melphalan (L-PAM)	Robins, 1997 (36)
		Ifosfamide (IFO)	Engelhardt, 1988 (34)
			Issels, 1990 (37)
			Westermann, 2003 (38)
Nitrosoureas	Impaired DNA repair	BCNU	Parks, 1983 (4)
	Improved pharmacokinetics		Bull, 1992 (39)
		Me-CCNU	Bull, 1992 (39)
Platinum agents	Impaired DNA repair	Cisplatin (CDDP)	Parks, 1983 (4)
	Altered plasma protein binding		Herman, 1982 (40)
			Engelhardt, 1990 (41)
			Robins, 1993 (42)
			Douwes, 2004 (43)
		Carboplatin (CBDCA)	Westermann, 2003 (38)
			Hegewisch-Becker, 2002 (44)
			Hegewisch-Becker, 2003 (45)
			Douwes, 2004 (43)
			Richel, 2004 (46)
			Strobl, 2004 (47)
		Oxaliplatin	Elias, 2004 (48)
			Hegewisch-Becker, 2002 (44)
Anthracyline antibiotics	Impaired DNA repair	Adriamycin	Engelhardt, 1990 (41)
	Enzyme activation		Bull, 2002 (49)
		Bleomycin	Herman, 1982 (40)
Antimetabolites	Increased drug transport	5-FU	Lange, 1983 (35)
	Cell cycle arrest		Larkin, 1979 (50)
	Impaired DNA repair		Bull, 2002 (49)
			Hegewisch-Becker, 2002 (44)
		Gemcitabine	Bull, 2004 (51)
Antiproliferatives	Impaired DNA repair	Etoposide (VP-16)	Barlogie, 1979 (52)
			Issels, 1990 (37)
			Westermann, 2003 (42)
Topoisomerase inhibitors	Impaired DNA repair	Irinotecan (CPT-11)	Hegewisch-Becker, 2003 (45)
			Elias, 2004 (48)
Taxanes	Microtubule disruption	Paclitacel	Strobl, 2004 (49)
	Apoptosis	Docetaxel	Strobl, 2004 (49)
Biological response modifiers	Increased anti-viral and antiproliferative activity	Interferon	Robins, 1989 (53)
			Bull, 2002, 2004 (34,49)

[a] References prior to 1980, or not in English, are not provided in the Bibliography at the end of this article.

given during the plateau phase of WBH, 10 min after target temperature was reached (56).

A major rationale for whole-body hyperthermia in cancer treatment is the ability to treat metastases, but this is actually a controversial issue. There have been no clinical studies specifically designed to study the effect of systemic hyperthermia on either the efficacy against metastatic disease or prevention of development of metastases. Increased survival in advanced malignancies is often interpreted to mean a reduction in metastatic disease, but direct measurement of the incidence and response of metastases is rare. Based on some animal studies, it has been suggested that systemic hyperthermia could actually promote the metastatic spread of tumor cells, but this has not been confirmed. One clinical study found an increase of tumor cells in blood 24 h after 41.8 °C WBH, but there was no evidence that this caused metastatic spread of disease (57). Several animal experiments do support the efficacy of whole-body hyperthermia against metastases. In mouse models of lung cancer and melanoma, the number of lung metastases was scored after repeated systemic microwave heating. It was found that the number of lung metastases was significantly reduced, and NK-cell activity was higher, in treated animals. The authors hypothesized that WBH interferes with the spread of organ metastases, possibly through a mechanism involving NK cells (13). Another study of mouse Lewis lung carcinoma in which the animals were treated with 60 min of systemic hyperthermia at 42 °C, demonstrated a reduction in the number and percentage of large metastases (>3 mm) on day 20 post-tumor implantation. Addition of radiation led to a reduction to 50% of control of the number of lung metastases as well as the percent of large metastases on day 20 (58). In a breast cancer ocult metastasis model in rats, 6 h of 40 °C whole-body hyperthermia combined with daily, low dose, metronomic irinotecan resulted in delayed onset, and reduced incidence, of axillary lymph node metastases compared to control in rats, as did treatment with 40 °C WBH alone. The combination therapy also reduced axillary metastasis volume. Interestingly, none of the therapies significantly affected inguinal lymph node metastases, but lung metastases were decreased in both the combination therapy and WBH alone groups. Rats treated with fever-range whole-body hyperthermia and metronomic irinotecan also survived significantly longer (36%) than control animals (59).

Systemic Hyperthermia and Radiotherapy. The augmentation of ionizing radiation induced tumor kill by hyperthermia is well documented for local hyperthermia and has led to numerous protocols combining whole-body hyperthermia with radiation therapy (60,61). Hyperthermia is complementary to radiation in several regards: ionizing radiation acts predominantly in the M and G_1 phases of the cell cycle while hyperthermia acts largely in S phase; radiation is most effective in alkaline tissues whereas hyperthermic cytotoxicity is enhanced under acidic conditions; radiation is not effective in hypoxic regions yet hyperthermia is most toxic to hypoxic cells. Thus when hyperthermia is combined with radiotherapy, both the hypoxic, low pH core of the tumor is treated as well as the relatively well perfused outer layers of the tumor. Furthermore, because of its vascular effects, hyperthermia enhances tumor oxygenation thus potentiating radiation cell kill. Hyperthermia also increases the production of oxygen radicals by radiation, and reduces the repair of DNA damage caused by ionizing radiation. Thus hyperthermia and radiotherapy together often have a synergistic effect, and this combination is now well accepted for treatment of a number of tumors.

Fever-Range WBH. Like systemic hyperthermia alone, combined modality treatments were initially aimed to achieve maximally tolerated temperatures. Such regimens, however, carry significant risk to the patient, require general anesthesia, and necessitate experienced, specialist personnel to provide careful monitoring of vital signs and patient care during the treatment. More recently, it has been appreciated that lower core body temperatures (39–40 °C) maintained for a longer time (4–8 h), much like fever, can indirectly result in tumor regression through effects on tumor vasculature, the immune response, and waste removal (detoxification). The optimum duration and frequency of mild hyperthermia treatment has, however, not yet been determined. Protocols range from single treatments of 4–6 h, or similar long duration treatments given once during each cycle of chemotherapy, to daily treatments of only 1 h. Several studies of mild, fever-range, whole-body hyperthermia with chemotherapy have demonstrated efficacy against a broad range of cancers (34,17) and clinical trials are currently being conducted at the University of Texas Health Science Center at Houston, Roswell Park Cancer Institute, New York, and by the German Interdisciplinary Working Group on Hyperthermia (62).

Systemic Hyperthermia and Metabolic Therapy. Increased rates of metabolic reactions lead to rapid turnover of metabolites, causing cellular energy depletion, acidosis, and consequent metabolic disregulation. Tumors, which have increased metabolic rates [glucose, adenomine triphosphate (ATP)] compared to normal cells, may be particularly sensitive to thermally induced energy depletion and this has been exploited in the Cancer Multistep Therapy developed by von Ardenne, which is a combined hyperthermia–chemotherapy–metabolic therapy approach to cancer (63). The core of this approach is systemic hyperthermia at 40–42 °C, sometimes with added local hyperthermia to achieve high temperatures within the tumor. A 10% solution of glucose is infused into the patient to achieve a high accumulation of lactic acid within the tumor that cannot be cleared because of sluggish blood flow and confers an increased sensitivity to heat to the tumor cells. Administration of oxygen increases the arterial oxygen pressure and stimulates lysozymal cytolysis. Finally low dose chemotherapy is added.

Palliation. Pain relief is reported by many patients receiving systemic hyperthermia treatment, whether with chemotherapy or radiation. Indeed, almost all patients undergoing thermoradiotherapy report pain relief. Immediate pain relief following treatment is likely to stem from an increased level of circulating β-endorphins, while longer term pain relief may be due to increased blood flow, direct neurological action, and disease resolution, for example, tumor regression in cancer patients, or detoxification. Meaningful improvements in quality of life typically result from such pain relief. Localized infrared therapy using lamps radiating at 2–25 μm is used for the treatment and relief of pain in numerous medical institutes in China and Japan.

Diseases Other than Cancer. Therapeutic use of heat lamps emitting IR radiation is commonplace throughout the Orient for rheumatic, neurological and musculoskeletal conditions, as well as skin diseases, wound healing, and burns. The improvements reported appear to be largely due to increased blood flow bringing nutrients to areas of ischemia or blood vessel damage, and removing waste products. Scientific reports of these treatments are, however, difficult to find. Application of heat via hot baths or ultrasound has long been standard in physical therapy for arthritis and musculoskeletal conditions, though ice packs are also used to counter inflammatory responses. Heat decreases stiffness in tendons and ligaments, relaxes the muscles, decreases muscle spasm, and lessens pain. Unfortunately, few clinical trials of efficacy have been performed, and methodological differences or lack of rigor in the studies hinder comparisons (64). A clinical trial in Japan reported a supposedly successful solution for seven out of seven cases of rheumatoid arthritis treated with whole-body IR therapy, and it is reported that the King of Belgium was cured of his rheumatoid arthritis in three months due IR treatments. Systemic hyperthermia with whole-body radiant heat units is being carried out in clinical centers as well as many private clinics in Germany for the purpose of alleviating rheumatoid arthritis. It has been proposed that the induction of TNF receptors by WBH may induce a remission in patients with active rheumatoid arthritis. The use of heat packs has long been standard to relieve the pain of fibromyalgia. Again, the therapeutic effect is believed to be due to increased circulation flushing out toxins and speeding the healing process. Whole-body hyperthermia treatment for fibromyalgia and chronic fatigue syndrome (CFS) is to be found in a number of private clinics. Hyperthermia increases the number and activity of white blood cells, stimulating the depressed immune system of the CFS patient.

Because of its immune stimulating effects, whole-body hyperthermia is a strong candidate for treatment of chronic progressive viral infections, such as HIV and hepatitis C. A clinical trial at the University Medical Center Utrecht, The Netherlands has evaluated extracorporeal heating to induce systemic hyperthermia of 41.8 °C for 120 min under propofol anesthesia for treatment of hepatitis C (65). Human immunodeficiency virus (HIV)-infected T cells are more sensitive to heat than healthy lymphocytes, and susceptibility increases when the cells are presensitized by exposure to tumor necrosis factor. Thus, induction of whole-body hyperthermia or hyperthermia specifically limited to tissues having a high viral load is a potential antiviral therapy for acquired immunodeficiency syndrome (AIDS). An Italian study has found treatment of AIDS with beta-carotene and hyperthermia to be synergistic, preventing progression of early disease and also increasing the survival time in patients with severe AIDS. A single treatment of low flow extracorporeal hyperthermia was found effective against AIDS associated Kaposi's sarcoma, though there was significant toxicity. Core temperature was raised to 42 °C and held for 1 h with extracorporeal perfusion and *ex vivo* blood heating to 49 °C. Complete or partial regressions were seen in 20/29 of those treated at 30 days post-treatment, with regressions persisting in 14/29 of those treated at 120 days post-treatment. At 360 days, 4/29 maintained tumor regressions with 1 patient being in complete remission still at 26 months (66).

THE FUTURE OF SYSTEMIC HYPERTHERMIA

While there is a resurgence of interest in systemic hyperthermia, this modality has not yet been adopted as a mainstream therapy, and optimal clinical trials have not yet been carried out. Well-designed, well-controlled, multicenter clinical trials need to be conducted. In order to unequivocally demonstrate the utility of whole-body hyperthermia in the treatment of cancer as well as other diseases, it will be important to accrue a sufficiently large number of patients who do not have end-stage disease. Thanks to the commercial availability of systemic hyperthermia systems, the variability between induction techniques at different institutions can be removed. Newer instrumentation, particularly near-IR radiant heat devices, along with treatment at lower temperatures (fever-range thermal therapy) should lead to significantly reduced toxicity. Better exploitation of the narrow window of effective temperatures within which the cellular effects of heat can be exploited yet damage remains minimal, and improved understanding of the biological interactions *in vivo* of systemic heat with chemotherapy and radiation will be essential to optimize therapy.

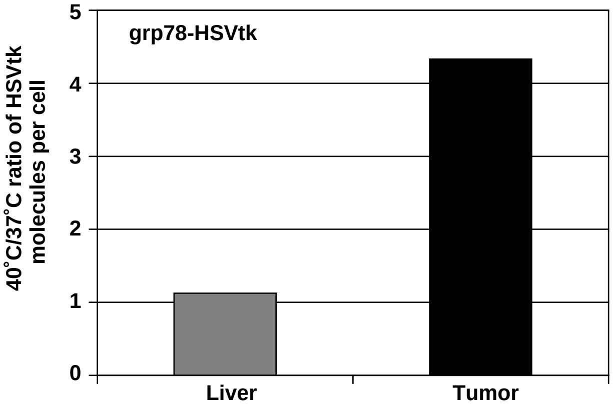


Figure 5. Fever range whole-body hyperthermia increases therapeutic gene (grp78-HSVtk) delivery in tumor.

The effects of systemic hyperthermia on tumor blood flow and vascular permeability have the potential to increase delivery of various small molecules, nanoparticles, and gene therapy vectors to tumors. Ferromagnetic nanoparticles can be heated by external magnetic fields and offer the potential for internal hyperthermia, both locally and systemically. Thermally sensitive liposomes that release their contents at designated temperatures are also of interest. The ability of systemic hyperthermia to aid in systemic delivery of gene therapy vectors (the holy grail of gene therapy) and enhance transfection of cells with therapeutic gene plasmids is under investigation in several laboratories (67,68), and shows potential along with targeted gene therapy via the heat shock response. For example, Fig. 5 shows a fourfold hyperthermic increase of therapeutic gene delivery to tumor when plasmid DNA was injected intravenously into mammary carcinoma bearing rats immediately after 6 h of whole-body hyperthermia at 40 °C. Thus systemic hyperthermia is likely to see increasing application as an enhancer of drug delivery.

There is a great deal of interest in the immunological consequences of whole-body hyperthermia, and as they become better understood, the combination of systemic hyperthermia with specific immunotherapies will undoubtedly be pioneered, not just for cancer but also, by analogy with fever, in a broad range of diseases.

SUMMARY

Systemic hyperthermia is founded on solid physical and biological principles and shows promise in the treatment of a number of diseases. Modern whole-body hyperthermia devices use IR-A radiation sources together with effective heat loss techniques to achieve a controlled, uniform temperature distribution throughout the body with minimal patient toxicity. A shift in paradigm has occurred away from achieving direct cell killing with short bouts of maximally tolerated temperatures, to inducing indirect curative effects through longer duration treatments at lower temperatures, and synergy with other modalities, such as radiotherapy. Better understanding of the interactions of elevated temperature with metabolic and genetic pathways will allow thermally driven targeted therapies. Of particular promise is the use of systemic hyperthermia as an immune system stimulator and adjunct to immunotherapy. Application of systemic hyperthermia to nanoparticle delivery and gene therapy is emerging. Whole-body hyperthermia is moving from being dubbed an alternative therapy to becoming a standard treatment and clinical hyperthermia centers are to be found all over the world.

Useful Websites

http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=hstat6.section.40680	Techniques and Devices Used to Produce Hyperthermia
http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=cmed.section.7813	Physics and Physiology of Heating
http://www.duke.edu/~dr3/hyperthermia_general.html	Clinical Hyperthermia Background
http://www.eurekah.com/isbn.php?isbn=1-58706-248-8&bookid=143&catid=50	Online book: *Locoregional Radiofrequency-perfusional and Whole Body Hyperthermia in Cancer Treatment: New Clinical Aspects*, E.F. Baronzio and A. Gramaglia (eds.), Eurekah Bioscience Database
http://www.esho.info/professionals/	European Society for Hyperthermic Oncology
http://www.hyperthermie.org/index2.html	German Interdisciplinary Working group on hyperthermia
http://www.uth.tmc.edu/thermaltherapy/	Systemic Thermal Therapy at the University of Texas

BIBLIOGRAPHY

1. Nauts HC. Bacterial pyrogens: beneficial effects on cancer patients. In: Gautherie M, Albert E, editors. Biomedical Thermology, Progress in Clinical Biological Research. New York: Alan R. Liss; 1982. p 687–696.
2. Law HT, Pettigrew RT. Heat transfer in whole body hyperthermia. Ann NY Acad Sci 1980;335:298–310.
3. Robins HI, et al. A non-toxic system for 41.8 °C whole body hyperthermia: results of a phase I study using a radiant heat device. Cancer Res 1985;45:3937–3944.
4. Parks LC, Smith GV. Systemic hyperthermia by extracorporeal induction techniques and results. In: Storm FK, editor. Hyperthermia in Cancer Therapy. Boston: Hall; 1983. pp 407–446.
5. van der Zee J, van Rhoon GC, Faithfull NS, van den Berg AP. Clinical hyperthermic practice: whole body hyperthermia. In: Field SB, Hand JW, editors. An Introduction to the Practical Aspects of Clinical Hyperthermia. London: Taylor & Francis; 1990.
6. Wust P, et al. Feasibility and analysis of thermal parameters for the whole-body-hyperthermia system IRATHERM-2000. Int J Hyperthermia 2000 Jul–Aug; 16(4):325–339.
7. Blazickova S, Rovensky J, Koska J, Vigas M. Effect of hyperthermic water bath on parameters of cellular immunity. Int J Clin Pharmacol Res 2000;20(1–2):41–46.
8. Vaupel P, Kallinowski F. Physiological effects of hyperthermia. In: Streffer C, editor. Hyperthermia and the Therapy of Malignant Tumors. Berlin/Heidelberg, Germany: Springer-Verlag; 1987.
9. Dudar TE, Jain RK. Differential response of normal and tumor microcirculation to hyperthermia. Cancer Res 1984;44(2):605–612.
10. Kong G, Braun RD, Dewhirst MW. Characterization of the effect of hyperthermia on nanoparticle extravasation from tumor vasculature. Cancer Res 2001;61:3027–3032.
11. Jain RK. Normalization of tumor vasculature: an emerging concept in antiangiogenic therapy. Science 2005;307(5706): 58–62.
12. Urano M, et al. Effect of whole body hyperthermia on cell survival, metastasis frequency, and host immunity in moderately and weakly immunogenic murine tumors. Cancer Res 1983;43:1039–1043.
13. Shen RN, et al. Whole-body hyperthermia decreases lung metastases in lung tumor-bearing mice, possibly via a mechanism involving natural killer cells. J Clin Immunol 1987;7(3):246–253.
14. Roigas J, Wallen ES, Loening SA, Moseley PL. Heat shock protein (HSP72) surface expression enhances the lysis of a human renal cell carcinoma by IL-2 stimulated NK cells. Adv Exp Med Biol 1998;451:225–229.
15. Burd R, et al. Tumor cell apoptosis, lymphocyte recruitment and tumor vascular changes are induced by low temperature, long duration (fever-like) whole body hyperthermia. J Cell Physiol 1998;177(1):137–147.
16. Ostberg JR, Gellin C, Patel R, Repasky EA. Regulatory potential of fever-range whole body hyperthermia on Langerhans cells and lymphocytes in an antigen-dependent cellular immune response. J Immunol 2001;167(5):2666–2670.
17. Kraybill WG, et al. A phase I study of low temperature, long duration (fever-range) whole body hyperthermia (WBH) in patients with advanced solid tumors: correlation with mouse models. Int J Hyperthermia 2002;18:253–266.
18. Atanackovic D, et al. 41.8 degrees C whole body hyperthermia as an adjunct to chemotherapy induces prolonged T cell activation in patients with various malignant diseases. Cancer Immunol Immunother Epub 2002 Oct 18 2002;51(11–12):603–613.
19. Barni S, et al. Lysosomal exocytosis induced by hyperthermia: a new model of cancer death. III. effect on liver metastasis. Biomed Pharmacother 1996;50:79–84.
20. Hou K. The level changes of peripheral blood PGE$_2$ and clinical significance in patients with malignant tumor before and after whole body hyperthermia. Proceedings of the 26th Congress of the International Clinical Hyperthermia Society, September 9–12, 2004, Shenzhen, China, 2004; p 66.
21. Manjili MH, et al. Subjeck, Cancer immunotherapy: stress proteins and hyperthermia. Int J Hyperthermia 2002;18(6): 506–520.
22. Ialenti A, et al. Inhibition of cyclooxygenase-2 gene expression by the heat shock response in J774 murine macrophages. Eur J Pharmacol 2005;509(2–3):89–96.

23. Repasky EA, Issels R. Physiological consequences of hyperthermia: Heat, heat shock proteins and the immune response. Int J Hyperthermia 2002;18:486–489.
24. Ostberg JR, Repasky EA. Emerging evidence indicates that physiologically relevant thermal stress regulates dendritic cell function. Cancer Immunol Immunother [Epub ahead of print]; Apr 28, 2005.
25. Berry JM, Michalsen A, Nagle V, Bull JM. The use of esmolol in whole-body hyperthermia: cardiovascular effects. Int J Hyperthermia 1997;13(3):261–268.
26. Robins HI, et al. Adjunctive therapy (whole body hyperthermia versus Ionidamine) to total body irradiation for the treatment of favorable B-cell neoplasms: a report of two pilot clinical trials and laboratory investigations. Int J Radiat Onco Biophys 1990;18:909–920.
27. Ohno S, et al. Haematological toxicity of carboplatin and cisplatin combined with whole body hyperthermia in rats. Br J Cancer 1993;68:469–474.
28. Haveman J, et al. Effects of hyperthermia on the peripheral nervous system: a review. Int J Hyperthermia 2004;20(4):371–391.
29. Calvert AH, et al. Early clinical studies with cis-diammine-1,1-cyclobutane dicarboylate platinum II. Cancer Chemother Pharmacol 1982;9:140–147.
30. Sapareto SA, Dewey WC. Thermal dose determination in cancer therapy. Int J Radiat Oncol Biol Phys 1984;10(6):787–800.
31. Thrall DE, et al. Using units of CEM 43 degrees C T90, local hyperthermia thermal dose can be delivered as prescribed. Int J Hyperthermia 2000;16(5):415–428.
32. Webb AG. Temperature measurement using nuclear magnetic resonance. Ann Reports NMR Spectrosc 2001;45:1–67.
33. Stafford RJ, Hazle JD, Glover GH. Monitoring of high-intensity focused ultrasound-induced temperature changes *in vitro* using an interleaved spiral acquisition. Magn Reson Med 2000;43(6):909–912.
34. Engelhardt R. Summary of recent clinical experience in whole-body hyperthermia combined with chemotherapy. Recent Results Cancer Res 1988;107:200–224.
35. Lange J, Zanker KS, Siewert JR, Eisler K, Landauer B, Kolb E, Blumel G. and Remy, W. Extracorporeally induced whole-body hyperthermia in conventionally incurable malignant tumor patients Med Wochenschr. 1983;108(13):504–509.
36. Robins HI, et al. Phase I clinical trial of melphalan and 41.8 degrees C whole-body hyperthermia in cancer patients. J Clin Oncol 1997;15:158–164.
37. Issels RD, Wilmanns W, editors. Recent Results in Cancer Research, Vol. 107: Application of Hyperthermia in the Treatment of Cancer. Berlin/Heidelberg: Springer Verlag; 1988.
38. Westermann AM, et al. Systemic Hyperthermia Oncologic Working Group, A Systemic Hyperthermia Oncologic Working Group trial, Ifosfamide, carboplatin, and etoposide combined with 41.8 degrees C whole-body hyperthermia for metastatic soft tissue sarcoma. Oncology 2003;64(4):312–321.
39. Bull JM, et al. Chemotherapy resistant sarcoma treated with whole body hyperthermia (WBH) combined with 1-3-bis(2-chloroethyl)-1-nitrosourea (BCNU). Int J Hyperthermia 1992; 8(3):297–304.
40. Herman TS, Sweets CC, White DM, Gerner EW. Effect of heating on lethality due to hyperthermia and selected chemotherapeutic drugs. J Natl Cancer Inst 1982;68(3):487–491.
41. Engelhardt R, et al. Treatment of disseminated malignant melanoma with cisplatin in combination with whole-body hyperthermia and doxorubicin. Int J Hyperthermia 1990;6(3): 511–515.
42. Robins HI, et al. Phase I clinical trial of carboplatin and 41.8 degrees C whole-body hyperthermia in cancer patients. J Clin Oncol 1993;11:1787–1794.
43. Douwes F, et al. Whole-body hyperthermia in combination with platinum-containing drugs in patients with recurrent ovarian cancer. Int J Clin Oncol 2004;9(2):85–91.
44. Hegewisch-Becker S, et al. Whole-body hyperthermia (41.8 °C) combined with bimonthly oxaliplatin, high-dose leucovorin and 5-fluorouracil 48-hour continuous infusion in pretreated metastatic colorectal cancer: a phase II study. Ann Onc 2002;13(8):1197–1204.
45. Hegewisch-Becker S, et al. Whole body hyperthermia (WBH, 41.8 °C) combined with carboplatin and etoposide in advanced biliary tract cancer. Proc Am Soc Clin Oncol 2003; (abstr. 1247), 22:311.
46. Richel O, et al. Phase II study of carboplatin and whole body hyperthermia (WBH) in recurrent and metastatic cervical cancer. Gynecol Oncol 2004;95(3):680–685.
47. Strobl B, et al. Whole body hyperthermia combined with carboplatin/paclitaxel in patients with ovarian carcinoma–Phase-II-study. J Clin Oncol -Proc Am Soc Clin Oncol 2004;22(14S).
48. Elias D, et al. Heated intra-operative intraperitoneal oxaliplatin plus irinotecan after complete resection of peritoneal carcinomatosis: pharmacokinetics, tissue distribution and tolerance. Ann Oncol 2004;15:1558–1565.
49. Bull JM, et al. Phase I study of long-duration, low-temperature whole-body hyperthermia (LL-WBH) with liposomal doxorubicin (Doxil), 5-fluorouracil (5-FU), & interferon-α (IFN-α), Proc Amer. Soc Clin Oncol 2002;(Abst. 2126).
50. Larkin JM, A clinical investigation of total-body hyperthermia as cancer therapy. Cancer Res, 1979;39(6 Pt 2):2252–2254.
51. Bull JM, et al. Update of a phase I clinical trial using fever-range whole-body hyperthermia (FR-WBH) + cisplatin (CIS) + gemcitabine (GEM) + metronomic, low-dose interferon-alpha (IFN-alpha), Proceedings of the International Congress on Hyperthermic Oncology, 20.-24.04., St. Louis, (Session 21 and Poster 689); 2004.
52. Barlogie B, Corry PM, Lip E, Lippman L, Johnston DA, Tenczynski TF, Reilly E, Lawson R, Dosik G, Rigor B, Hankenson R, Freireich EJ Total-body hyperthermia with and without chemotherapy for advanced human neoplasms. Cancer Res 1979;39(5):1481–1489.
53. Robins HI, et al. Phase I trial of human lymphoblastoid interferon with whole body hyperthermia in advanced cancer. Cancer Res 1989;49(6):1609–1615.
54. Gerad H, van Echo DA, Whitacre M, Ashman M, Helrich M, Foy J, Ostrow S, Wiernik PH, Aisner J. Doxorubicin, cyclophosphamide, and whole body hyperthermia for treatment of advanced soft tissue sarcoma. Cancer. 1984 Jun 15;53(12):2585–91.
55. Baba H, et al. Increased therapeutic gain of combined cis-diamminedichloroplatinum (II) and whole body hyperthermia therapy by optimal heat/drug scheduling. Cancer Res 1989;49(24 Pt. 1):7041–7044.
56. Katschinski DM, et al. Optimization of chemotherapy administration for clinical 41.8 degrees C whole body hyperthermia. Cancer Lett 1997;115(2):195–199.
57. Hegewisch-Becker S, et al. Effects of whole body hyperthermia (41.8 degrees C) on the frequency of tumor cells in the peripheral blood of patients with advanced malignancies. Clin Cancer Res 2003;9(6):2079–2084.
58. Teicher BA, Holden SA, Ara G, Menon K. Whole-body hyperthermia and lonidamine as adjuvant therapy to treatment with cisplatin with or without local radiation in mouse bearing the Lewis lung carcinoma. Int J Hyperthermia 1995;11(5):637–645.
59. Sumiyoshi K, Strebel FR, Rowe RW, Bull JM. The effect of whole-body hyperthermia combined with 'metronomic' chemotherapy on rat mammary adenocarcinoma metastases. Int J Hyperthermia 2003;19(2):103–118.
60. Overgaard J, et al. Randomised trial of hyperthermia as adjuvant to radiotherapy for recurrent or metastatic

malignant melanoma, European Society for Hyperthermic Oncology. Lancet 1995;345(8949):540–543.

61. Hehr T, Wust P, Bamberg M, Budach W. Current and potential role of thermoradiotherapy for solid tumours. Onkologie 2003;26(3):295–302.
62. Hildebrandt B, et al. Current status of radiant whole-body hyperthermia at temperatures > 41.5 degrees C and practical guidelines for the treatment of adults. The German Interdisciplinary Working Group on Hyperthermia. Int J Hyperthermia 2005;21(2):169–183.
63. Hildebrandt B, et al. Whole-body hyperthermia in the scope of von Ardenne's systemic cancer multistep therapy (sCMT) combined with chemotherapy in patients with metastatic colorectal cancer: a phase I/II study. Int J Hyperthermia 2004;20(3):317–333.
64. Robinson V, et al. Thermotherapy for treating rheumatoid arthritis. Cochrane Database Syst Rev 1: CD002826; 2002.
65. van Soest H, van Hattum J. New treatment options for chronic hepatitis C. Adv Exp Med Biol 2003;531:219–226.
66. Pontiggia P, Rotella GB, Sabato A, Curto FD. Therapeutic hyperthermia in cancer and AIDS: an updated survey. J Environ Pathol Toxicol Oncol 1996;15(2–4):289–297.
67. Li CY, Dewhirst MW. Hyperthermia-regulated immunogene therapy. Int J Hyperthermia 2002;18(6):586–596.
68. Okita A, et al. Efficiency of lipofection combined with hyperthermia in Lewis lung carcinoma cells and a rodent pleural dissemination model of lung carcinoma. Oncol Rep 2004;11:1313–1318.

Further Reading

Bakhshandeh A, et al. Year 2000 guidelines for clinical practice of whole body hyperthermia combined with cytotoxic drugs from the University of Lübeck and the University of Wisconsin. *J Oncol Pharm Practice* 1999;5(3):131–134.

Field SB, Hand JW, editors. An Introduction to the Practical Aspects of Clinical Hyperthermia. London: Taylor & Francis; 1990.

Gautherie M, editor. Methods of External Hyperthermic Heating. Berlin/Heidelberg: Springer Verlag; 1990.

Gautherie M, editor. Whole Body Hyperthermia: Biological and Clinical Aspects. Berlin/Heidelberg: Springer Verlag; 1992.

Hahn GM. Hyperthermia and Cancer. New York: Plenum Press, 1982.

Hildebrandt B, et al. Current status of radiant whole-body hyperthermia at temperatures > 41.5 degrees C and practical guidelines for the treatment of adults. The German Interdisciplinary Working Group on Hyperthermia. *Int J Hyperthermia* 2005;21(2):169–183.

Issels RD, Wilmanns W, editors. Recent Results in Cancer Research, Vol. 107: Application of Hyperthermia in the Treatment of Cancer. Berlin/Heidelberg: Springer Verlag; 1988.

Nussbaum GH, editor. Physical Aspects of Hyperthermia. American Association of Physicists in Medicine Medical Physics Monograph No. 8. New York: American Institute of Physics; 1982.

Guan J, et al. The clinical study of whole-body hyperthermia (WBH) improving the survival state of patients with advanced cancer. Proc 26th Congress of the International Clinical Hyperthermia Society, Sept. 9–12, 2004, Shenzhen, China; 2004; p 66.

Kurpeshev OK, Tsyb AF, Mardynsky YS. Whole-body hyperthermia for treatment of patients with disseminated tumors- Phase II. In: P.H. Rehak, K.H. Tscheliessnigg, editors. Proceedings 22nd. Annual Meeting of the European Society for Hyperthermic Oncology, June 8–11, 2005, Graz, Austria, 2005; p 103.

Hou K. Assessment of the effects and clinical safety of the treatment of advanced malignant tumor with extracorporeal whole body hyperthermia. Proceedings of the 26th Congress of the International Clinical Hyperthermia Society, Sept. 9–12, 2004, Shenzhen, China; 2004 p 71.

See also BIOHEAT TRANSFER; HEAT AND COLD, THERAPEUTIC; HYPERTHERMIA, INTERSTITIAL; HYPERTHERMIA, ULTRASONIC; RADIATION DOSIMETRY FOR ONCOLOGY.

HYPERTHERMIA, ULTRASONIC

DIMPI PATEL
DHANUNJAYA LAKKIREDDY
ANDREA NATALE
The Cleveland Clinic Foundation
Cleveland, Ohio

INTRODUCTION

The use of elevated temperature as a form of medical treatment has been fairly ubiquitous across cultures throughout the course of time. The earliest record of heat for therapeutic use was found in an Egyptian surgical papyrus dated to 3000 BC (1). Hippocrates, considered by many to be the father of medicine, used heat to treat breast cancer. He based his practice of medicine on an ancient Greek ideology that advises using heat after trials of surgery and medications have failed (2). German physicians in the 1800s noted cases where cancer patients had developed high fevers secondary to infections that resulted in a miraculous disappearance of their tumors (3). These observations provided inspiration for the development of several techniques that attempted to induce hyperthermia. One such popular method entailed wrapping a patient's body in plastic and then dipping him in hot wax. Another popular technique involved removing a portion of the patient's blood, heating it, and then transfusing the warmed blood back to the patient's body, thereby creating systemic hyperthermia (4). These treatments had varied success rates, often culminating in fatality, and were subsequently discarded. Thus, the interest in hyperthermia lessened in the face of more conventional cancer treatments (e.g., chemotherapy and radiation). The current revival of interest in hyperthermia has resulted from a combination of clinicians searching for a therapeutic mode other than chemotherapy and radiation, in tandem with several preliminary randomized clinical trials in a small selected group of patients that have shown marked improvement in disease states with the use of either hyperthermia alone or particularly as an adjuvant to other more traditional modalities.

Traditionally, conventional hyperthermia has been defined as a therapeutic elevation of whole body temperature or target tissue while maintaining low enough temperatures to avoid tissue coagulation (3). This definition of hyperthermia can be broadened to include the therapeutic elevation of temperature to cause tissue destruction and coagulation, such as that implemented in HIFU (high intensity focus ultrasound) procedures. Classically, microwaves, radio frequency (RF), electromagnetic radiations, or ultrasounds have been used to heat tissue to 40–44 °C (5). This article compares and contrasts electromagnetic waves to ultrasonic waves as a heating modality, explain the physics behind ultrasound generation, and explores the thermal and mechanical biophysics involved with ultrasound delivery to tissue. Then, the medical fields that are currently benefitting from conventional ultrasound hyperthermia and HIFU are considered, and finally some of the many applicators involved with thermal ultrasound delivery are evaluated.

ULTRASOUND VERSUS ELECTROMAGNETIC RADIATION

Electromagnetic waves were often used in various applications of conventional hyperthermia treatments. However, ultrasound has emerged as a better option because of its shorter wavelength and lower energy absorption rate, which make it easier to control and to localize the area that is being heated. For example, for a half-power penetration depth of 4 cm, the ultrasound wavelength in tissues (e.g., muscle) is 1 mm; however, electromagnetic wavelength required for the same transmission is 500 mm. Focusing energy into a volume smaller than a wavelength is generally not possible. Using 500 mm (~40 MHz) of electromagnetic waves to heat a tumor that is situated 4 cm below the skin with proportions of 6 cm in diameter in the upper abdomen results in a temperature elevation of the entire abdomen including the spleen, liver, and all major vessels. More than one-half of the body's cardiac output circulates through the abdominal area, and this widespread heating results in a systemic elevation of temperature, thereby limiting the use of electromagnetic radiation for tumors in the body cavity (3). Electromagnetic waves are currently limited to regional hyperthermia and treating very superficial tumors (6). Conversely, ultrasound that has a wavelength of 1 mm can be focused within the area of the tumor, thus allowing less energy to be radiated to other areas of the body, resulting in less damage to surrounding healthy tissue. The current fabrication technology allows for practical applicator dimensions and multiple transducer configurations that makes it possible to control and shape a wide variety of ultrasound beams. The use of focused transducers or electronically phased arrays allow for better localization and temperature control of the target tissue (7). In contrast to these positive attributes, high acoustic absorption at bone–soft tissue interface and reflection from gas surfaces may make certain therapeutic scenarios difficult.

GENERATION AND PROPAGATION OF ULTRASOUND

Ultrasonic Transducers

In order to generate ultrasonic waves for tissue warming, a transducer containing piezoelectric crystals is required. Piezoelectric crystals are found in Nature or can be artificially grown. Quartz and synthetic ferroelectric ceramics (e.g., lead metanobiate, lead zirconate, and titanates of barium) all have strong piezoelectric properties (8). The ceramic most commonly used in the fabrication of ultrasound transducers is synthetic plumbium zirconium titante (PZT). Transducers are manufactured by applying an external voltage to these ferroelectric materials to orient their internal dipole structure. They are then cooled to permanently maintain their dipole orientation. Finally, they are cut into any desired shape, such as spherical bowls for focused ultrasonic fields (3,8). Ultrasound transducers have electrodes attached to the front and back for application and detection of electrical fields. With the application of an alternating electrical field parallel to the surface of piezoelectric material, the crystals will contract and vibrate for a short time with their resonant frequency. The frequency at which the transducer is able to vibrate is indirectly proportional to its thickness; higher frequencies are a result of thinner transducers, lower frequencies a result of thicker transducers (8).

Piezoelectric crystals are able to contract or expand when an electrical field is applied to them because dipoles within the crystal lattice will realign themselves as a result of attractive and repulsive forces causing a change in physical dimension of the material in the order of nanometers (electrostriction or reverse piezoelectric effect). When echos are received, the ultrasound waves will compress and expand the crystals (8). This mechanical stress causes the dipoles to realign on the crystal surface creating a net charge (piezoelectric effect) (Fig. 1).

Transducers function optimally when there is broad bandwidth in the frequency domain and short impulse response in the time domain. Also, when there is little electroacoustical conversion inefficiency, and little mismatch between the electrical impedances of the generator and the transducer (3,9). A transducer's ability to transmit energy is dependent on the characteristics of acoustic impedances and its contact medium. Both the density of the material and propagation velocity of ultrasound waves will determine its impedance. When both impedances match, then less energy is lost through reflection back into the transducer. For example, at the interface between air and the transducer, most of the energy will be reflected back to the transducer and will travel to the opposite direction because air has ~16 times less impedance than the transducer. If the transducer is a half wavelength in thickness, the reflected wave arrives at the opposite surface in phase with the direction of its motion and can then be transmitted into the medium. Since the efficiency at which a transducer transmits energy has a direct relationship to the degree of impedance match, efficiency can be increased significantly by adding an impedance matching layer of a quarter wavelength thickness, subsequently making the characteristic impedance equal to the geometric average of those of the transducer and the loading medium (3,8).

RADIATION FIELD OF ULTRASONIC TRANSDUCERS

The radiation field of an ultrasonic transducer depends on its physical properties and the transmission characteristics of the medium through which it will pass. Conventional planar transducers create a nonfocused field, whereas some modifications to the same transducer can create a focused field.

NONFOCUSED FIELDS

Planar Transducers

Ultrasonic waves that are radiated from the transducer surface can be described as a tightly packed array of

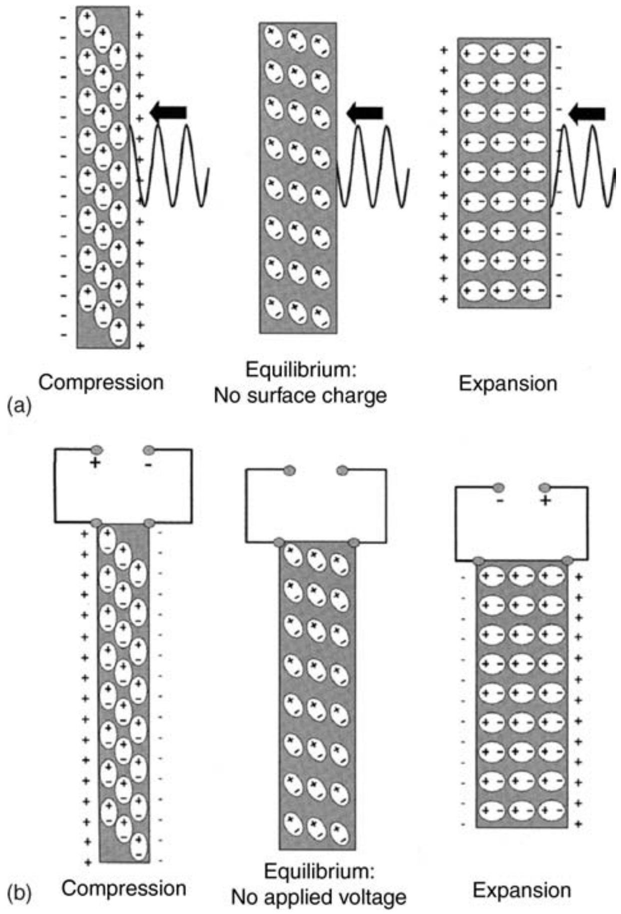


Figure 1. A piezoelectric compound consists of aligned molecular dipoles. (a) At equilibrium, the dipoles are arranged in a configuration that results in a neutral net charge. When the piezoelectric compound is mechanically stressed (e.g., an ultrasound wave) the element changes its physical dimensions. At peak pressure amplitudes, the element will contract. When no stress is placed upon the element it is in equilibrium. At peak rarefaction, the element will expand. This realignment of dipoles results in the production of a net positive or negative surface charge. (b) When an electrical field is applied to the piezoelectric element the dipoles can be realigned in response to attractive or repulsion forces. This rearrangement results in either expansion or contraction of the element. In the absence of an applied electrical field the element is in equilibrium and has a neutral net charge. (Published with the permission from Ref. 8).

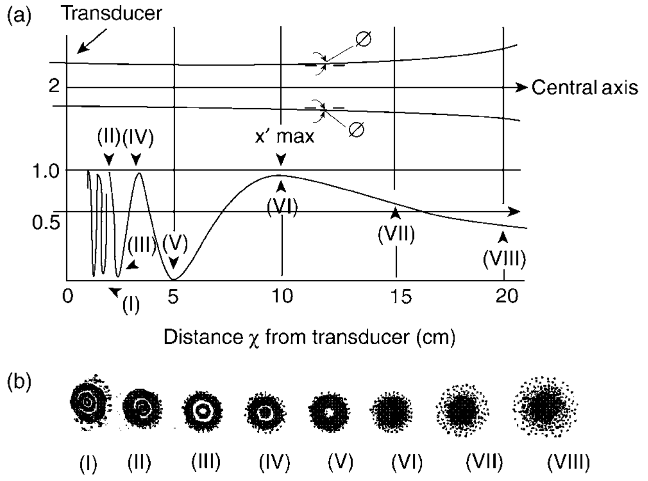


Figure 2. A planar transducer operating in a continuous wave mode. (a) The envelope containing almost all of the ultrasonic energy. (b) The relative intensity of the ultrasonic beam along a central axis. The intensity across the field fluctuates greatly at small distances from the surface of the transducer. At greater distances along the central axis the intensity distribution across the field stabilizes and deteriorates with beam divergence. (c) Ring diagrams illustrating the energy distribution at positions indicated in (b). In the near field, the ring beam pattern is collimated, but at greater distances from the transducer surface the beam diverges. (Published with permission from Ref. 3).

separate point sources of sound energy (Fig. 2a). Each of these points emits a spherical wavelet (Fig. 3). These waves interact both constructively and destructively creating a diffraction pattern. Any point in the medium is a compilation of all the sources that reach that target at that period of time. This diffraction pattern can be calculated using Huygen's principle. Two separate transducers whose emission fields interact in the same media are subject to the same laws of construction and destruction. Planar transducers operating in continuous wave mode are shown in (Fig. 2). In the Fresnel zone or the near field, the beam energy distribution is collimated, which is a result of the many destructive and constructive interactions of the spherical wavelets (Figs. 2c and 3). The beam path is a function of the dimension of the active part of the transducer surface, thus the beam diameter that is converging at the end of the near field is approximately one-half of the size of transducer diameter. The intensity and pressure amplitudes fluctuate greatly at small distances from the surface transducer (Fig. 2b). As the distance from the transducer surface increases, the beam path diverges (Fig. 2c and 3). In large diameter, high frequency transducers, there is less beam divergence in the far field. After a certain distance from the transducer surface, the intensity stabilizes; however, intensity along the axis deteriorates along with beam divergence (Fig. 2b). Circular, square, and rectangular transducers have similar fields; albeit, circular transducers have more pronounced fluctuations of intensity in the near field (3,8).

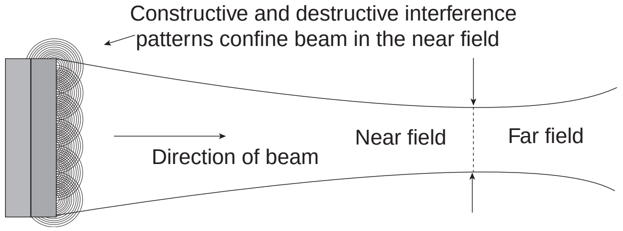

Constructive and destructive interference
patterns confine beam in the near field
Direction of beam
Near field
Far field

Figure 3. Ultrasonic waves that are radiated from the transducer surface are described as a tightly packed array of separate point sources. Each of these points radiates a spherical wavelet (Huygen's principle). These spherical wavelets will interact constructively and destructively. In the near field, these interactions result in a convergent beam pattern. In the far field, the beam pattern diverges. (Published with permission from Ref. 8).

FOCUSED FIELDS

Single-Element Transducers

When an ultrasonic wave travels through different media, the laws of geometric optics can be applied. Ultrasonic waves can be reflected, refracted, and scattered. When there is a high degree of impedance mismatch between the generator and transducer, ultrasonic waves will be reflected back into the transducer. The angle of reflection is equal to the angle of impedance, much like that of a mirror. Single element transducers can be focused by using a curved acoustic lens or a curved piezoelectric element (8). When an ultrasonic wave goes through two media with different propagation velocities there is a certain degree of refraction. Ultrasonic propagation through water is 1500 m·s^{-1}. In order to focus the ultrasound field, a lens of plastic (e.g., polystyrene), which has a higher propagation velocity, is placed between the transducer and the water media; these converging lenses are then concave to the water media and at plane with the transducer interface. In an unfocused transducer, the focal length is directly proportional to the transducer frequency and diameter. In a focused single element transducer, the focal distance is brought closer to the transducer surface. The focal distance is defined as the distance between the transducer surface and the portion of the beam that is narrowest. The focal zone, which is the area of best lateral resolution, is defined as the area at which the width of the beam is less than two times the width at the focal distance (3,8) (Fig. 4). The focal zone is dependent on the aperture and the wavelength of the ultrasound. The focal area through which 84% of the ultrasound passes is two to four wavelengths in hyperthermia systems. With ultrasonic transducers, the intensity distribution dimensions are a function of frequency and aperture. Therefore, the larger the aperture, the shorter the focal region, the higher the frequency and the smaller the diameter of the beam (Fig. 5). Ceramic curved bowl-shaped transducers, while more efficient than a lens, do not have the versatility of a lens. Once a ceramic bowl is fabricated, the focal length is set. Lenses can be interchanged creating a variety of focal lengths with one transducer (8).

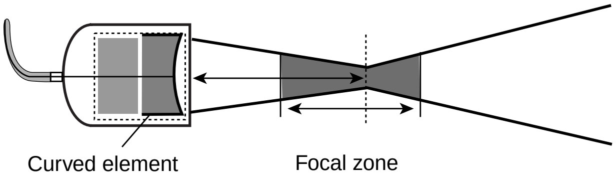


Figure 4. The focal zone is the area of optimal lateral resolution. The use of a curved element or an acoustic lens allows the focal distance to be brought closer to the transducer surface. The use of a curved element decreases the beam diameter at the focal distance and increases the angle of beam divergence far field. The focal zone, which is the are of best lateral resolution, is defined as the area at which the width of the beam is less than two times the width at the focal distance. (Published with permission from Ref. 8).

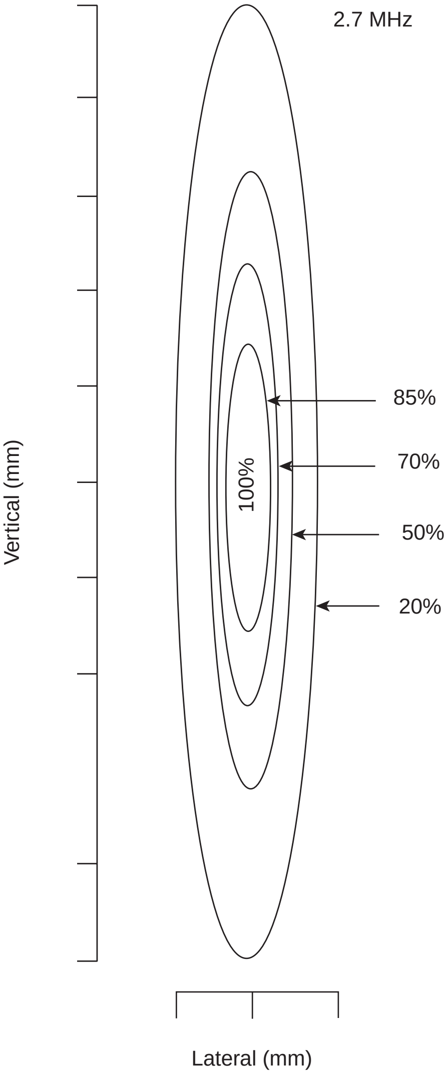


Figure 5. The intensity distribution in the focus of an ultrasound transducer. The diameter and the length are a function of frequency. The lower the frequency, the larger the diameter, and the smaller the aperture, the longer the focal region. (Published with permission from Ref. 3).

Multielement Transducers

Linear array transducers contain 256–512 narrow piezoelectric elements. They produce a beam by firing a portion of the total number of elements as a group. If a single element were fired the beam pattern would be divergent in the near field. By firing a group of elements, it is possible to focus and converge the beam. All the individual beams interact both constructively and destructively to produce a collimated beam. A phase array transducer is composed of 64–128 elements. The ultrasound beam can be steered and focused without moving the transducer by electrically activating the separate elements on the transducer surface at slightly different times (8).

Table 1. Acoustic Impedance[a]

Tissue	Z, rayls
Air	0.0004×10^6
Lung	0.18×10^6
Fat	1.34×10^6
water	1.48×10^6
Kidney	1.63×10^6
Blood	1.65×10^6
Liver	1.65×10^6
Muscle	1.71×10^6
Skull bone	7.8×10^6

[a] Measured in rayls. $z = pc$ (z = impedance, p = sound pressure, c = speed of sound), for air, water, and selected tissues. (Published with permission from Ref. (8).

Table 2. Density and Speed of Sound in Tissues and Materials for Medical Ultrasound.[a]

Material	Density, kg·m^{-3}	c, m·s^{-1}	c, mm·s^{-1}
Air	1.2	330	0.33
Lung	300	600	0.60
Fat	924	1450	1.45
water	1000	1480	1.48
Soft tissue	1050	1540	1.54
Kidney	1041	1565	1.57
Blood	1058	1560	1.56
Liver	1061	1555	1.55
Muscle	1068	1600	1.60
Skull bone	1912	4080	4.08
PZT[b]	7500	4000	4.00

[a] Published with permission from Ref. 8.

[b] PZT, lead cisconate .inanate

PROPAGATION OF ULTRASOUND IN BIOLOGICAL TISSUES

The journey of an ultrasound wave through human tissue is sometimes arduous. As the waves propagate through different biological media, they are subject to reflection, refraction, scattering, and absorption (8). When there is a difference in acoustic impedance between the boundaries of two tissues, reflection occurs (Table 1). There is 100% reflection at the air–skin interface. However, if a coupling medium (e.g., gel) is used, reflection is reduced to 0.1%. When the beam is not perpendicular to tissue boundary, the transmitted ultrasound energy undergoes a change in direction at the boundary; this directional change is termed refraction. As waves propagate through tissue they must overcome internal friction resulting in a loss of energy. The mechanical energy that is lost is converted to heat, which is termed absorption. At higher frequencies, ultrasonic waves move quicker, thus forcing the molecules to move against each other creating friction. The more these molecules move, the more energy is consumed or absorbed, and subsequently will be converted to heat. The speed at which the ultrasonic wave travels is dependent on the arrangement of the molecules. If they are densely arranged, they will collide sooner when they are exposed to a stimulus, and will lose energy quickly and at shorter distances. Ultrasonic waves can travel through the skin without much absorption until they reach tissues with high collagen content, (e.g., bone, periosteum, ligaments, capsules, fascia, and tendons). Ultrasonic waves travel through most solid tissues at a speed of 1500–1600 m·s^{-1}. Its velocity in fat and postmenopausal breast tissue may be as low as 1450 m·s^{-1} and the lens of the eye ~1600 m·s^{-1}. As a general rule, ultrasonic waves move through soft tissue with relatively little reflection or refraction (3,8) (Table 2).

Ultrasonic speed through bone is ~4080 m·s^{-1}. Bone easily absorbs ultrasonic energy and reflects it to tissues that are located at the bone–tissue interface. Since bone absorbs ultrasonic energy so readily, it heats up very quickly, consequently making it harder to control temperature. Thus, bone and its immediate surrounding tissue were once considered problematic for therapeutic use of ultrasonic hyperthermia (3,8). Nevertheless, in a recent study, Moros et al. noted that the presence of underlying bone in superficial unfocused ultrasound hyperthermia could actually be exploited to induce more uniform and enhanced temperature distributions in superficial target volumes. In particular, they have shown that the presence of bone in superficial target volumes enhances temperature elevation not only by additional direct power deposition from acoustic reflection, but also from thermal diffusion from the underlying bone (10).

The intensity at which an ultrasonic beam is transmitted has an effect on target tissue temperature. Intensity is defined as the amount of power per unit area. Doubling the amount of power used will result in quadrupling the intensity. Ultrasonic waves will lose intensity as they propagate further into the tissue. The attenuation coefficient is the relative intensity loss per centimeter of travel in a given medium (Table 3). Beam divergence, absorption, and scattering will also cause a loss in intensity of the propagating beam. The absorption coefficient of the tissue being exposed to us determines the target temperature that tissue will reach. The absorption coefficient is dependent on the density of the tissue and will linearly increase at higher frequencies. The absorption coefficient in soft tissue is 4–10 times lower than that of bone, and therefore bone heats more quickly (3). At short exposure times (e.g., < 0.1 s), temperature and intensity are directly propor-

Table 3. Attenuation Coefficients for Selected Tissues at 1 MHz.[a]

Tissue Composition	Attenuation Coefficient 1 MHz beam, dB·cm^{-1}
Water	0.0002
Blood	0.18
Soft tissues	0.3–0.8
Brain	0.3–0.5
Liver	0.4–0.7
Fat	0.5–1.8
Smooth muscle	0.2–0.6
Tendon	0.9–1.1
Bone, cortical	13–26
Lung	40

[a] Published with permission from Ref. 8.

tional. However, as the time intervals increase, other factors in addition to intensity, (e.g., blood perfusion), must be considered. An approximate estimate of the ultrasonic energy requirements for heating a target volume to therapeutic temperature depends on assessing thermophysical properties of that tissue, intensity of the ultrasound beam, ultrasonic absorption coefficient, and additional factors (e.g., blood circulation to target tissue) (3,8,11). The thermal index is defined as the ratio of acoustic power created by the transducer to raise the target area by 1 °C. This is calculated by using an algorithm that takes into account the ultrasonic frequency, beam area, and the acoustic power output of the transducer (8).

Ultrasonic waves act on tissues thermally and mechanically. Mechanical effects on tissues via ultrasound include acoustic torque, acoustic streaming, radiation force, stable cavitation, and unstable cavitation (11). Any object situated within an acoustic field will be subject to acoustic pressure and acoustic force. Acoustic pressure in a standing wave field is inversely proportional to velocity. Acoustic torque results from variations in the acoustic field, which can be described as a time-independent twisting action. Acoustic torque causes a rotational movement of cells and intracellular organelles in the medium. Acoustic streaming describes the movement of fluid in an ultrasonic field. The compression phase of an ultrasonic wave deforms tissue molecules. Radiation force affects gas bubbles that are in the tissue fluids. Negative pressure induces the bubbles originally dissolved in the medium to fall out of solution. With positive and negative pressure wave fluctuations, these bubbles expand and contract without reaching critical size (stable cavitation). Unstable cavitation occurs when bubbles collapse violently under pressure after growing to critical size due to excessive energy accumulation. This implosion produces large, brief local pressure and temperature release, as well as causing the release of free radicals. Organs that are air-filled, (e.g., the lungs or intestines), are subject to greater chance of unstable cavitation. Unstable cavitation is somewhat random, and as such it may lead to uncontrollable tissue destruction (8,11). Bubble growth can be limited by low intensity, high frequency, and pulsed ultrasound. Higher frequency means shorter cycle duration so time for bubble growth is regulated. Pulsed ultrasound restricts the number of successive growth cycles and allows the bubble to regain its initial size during the off period. The mechanical index estimates the possibility of cavitation occurrence. The mechanical index is directly proportional to peak rarefaction pressure, and inversely proportional to the square root of the ultrasound frequency (8).

MEDICAL APPLICATIONS OF CONVENTIONAL HYPERTHERMIA

Ultrasound as a heating modality has been used in several different medical fields. It is used in treating sprains, bursitis, joint inflammation, cardiac ablations, and in gynecology. However, the main area conventional hyperthermia is currently used is in oncology. The use of conventional hyperthermia as an oncologic treatment is supported by a plethora of studies that demonstrate that heat on cell lines and on animal tumor transplant models can result in tumor regression; however, it is rarely used alone because its efficacy is greatly potentiated in combination with radiation or chemotherapy. Conventional hyperthermia treatments elevate target tissue temperatures to 42–46 °C (12). Treatment times are usually between 30 and 60 min. Treatment applications are administered once or twice a week and are applied in conjunction with or not long after radiation. In all of the recent phase III trials, a sequential delivery scheme was used. This means that radiation and hyperthermia were administered separately, with radiation preceding hyperthermia treatments (13). Tumoricidal effects *in vivo* are achieved at temperatures between 40 and 44 °C (5). Large tumors often have an inadequate blood supply and resultantly, have difficulty meeting their requirements for oxygen and nutrients. This situation creates a hypoxic environment that is low in pH (2–3) (3,5,14). When tumor cells are heated to therapeutic temperatures, their cellular metabolic processes are accelerated, thereby further increasing the demands for oxygen and nutrients in an already depleted environment. Most tumor cells are unable to reproduce in this hostile environment, resulting in termination of tumor growth and shrinkage of the tumor (5,15). In temperatures > 40 °C, protein denaturation has been observed as the main mechanism of cellular death. Widespread protein denaturation results in structural changes in the cytoskeleton and the cell membrane, and in enzymes that are necessary for deoxyribonucleic acid (DNA) synthesis, cellular division, and cellular repair (5). Hyperthermic efficacy is a function of temperature and exposure time. To quantify, at temperatures > 42.5– 43 °C, the exposure time can be halved with each 1 ° temperature increase to give an equivalent cell kill (5,16). Healthy tissues remain undamaged at temperatures of 44 °C for a 1 h duration (5,17). The exceptions are central nervous tissues, which suffer irreversible damage after being exposed to heat at temperatures ranging from 42 to 42.5 °C for >40–60 min (5,18). Peripheral nervous tissue that has been treated for $>$ 30 min at 44 °C or an equivalent dose results in temporary functional loss that is reversed in 4 weeks (5,19). Therefore, since a small difference in temperature produces a large difference in the amount of cells killed, it is important to be able to have good control on the site and duration of heat delivery to reduce the damage to surrounding healthy tissue.

RADIATION COUPLED WITH CONVENTIONAL HYPERTHERMIA

While hyperthermia independently has been found to have antitumor effects, its efficacy is greatly potentiated when coupled with radiation. Cells that are in a low pH hypoxic environments, those that are in the S or M phases of cell division, and those that are malnourished are relatively resistant to radiation (5,7). Hyperthermia increases radiation damage and prevents cellular repair of damaged DNA (5,16). Hyperthermia increases blood perfusion via vasodilation which results in increased oxygenation, thus allowing increased radiosensitivity (5,7,16). Response rates with hyperthermia alone are ~15%, with radiotherapy ~35%, with combined radiotherapy, and hyperthermia ~70% (20). There have been many U.S. and European clinical trials that support substantial improvement in patients who have been treated with a combination of radiation and hyperthermia. Examples of some recent trials include randomized multiinstitutional phase III trials for treating melanoma (20,21), glioblastoma multiforme (20,22), chest wall recurrence of breast cancer (20,23), head and neck cancer (20,24,25), head and neck in superficial measurable tumors (20,26,27), in various recurrent persistent tumors (20,28), cervical cancer (29), uterine cancer (30) and in locally advanced pelvic tumors (20,31) (Table 4). Trial success rates were very dependent on the uniformity of temperature delivery. In the past, trials had often provided mediocre results because temperatures were ~1–1.5 °C too low and consequently not able achieve adequate tumoricidal levels (7). It is often difficult to uniformly heat larger tumor (3,7). When radiation and hyperthermia are used simultaneously excellent radiation delivery is achieved, often resulting in tumor regression; however, its delivery is equally as toxic to healthy cells that necessitate the need for a very precise delivery system.

CHEMOTHERAPY IN CONJUNCTION WITH CONVENTIONAL HYPERTHERMIA

Chemotherapeutic efficacy is enhanced by hyperthermia (5,20,34,35) (Table 5). As areas are heated, perfusion is increased, thus allowing an increase in drug concentrations in areas of the tumor that are poorly vascularized, increased intracellular drug uptake, and enhanced DNA damage. Drugs (e.g., mitomycin C, nitrosureas, cisplatin, doxorubicin, and mitoxantrone) are subject to less drug resistance when used with heat. The synergistic effect of chemotherapy and hyperthermia was demonstrated in virtually all cell lines treated at temperatures >40 °C for alkylating drugs, nitroureas, and platin analogs dependent on exposure time and temperature. Chemotherapeutic agents can be revved up 1.2–10 times with the addition of heat (5). *In vivo*, experiments showed improvement when doxorubicin and mitoxantrone were combined with hyperthermia. However, antimetabolites vinblastine, vincristine, and etoposide did not show improvement with the addition of hyperthermia. In animal studies, increased toxicities were seen in skin (cyclophosphamide, bleomycin), heart (doxorubicin), kidney (cisplatin, with a core temperature >41 °C), urinary tract (carmustine, with core temperatures >41 °C), and bone marrow (alkylating agents and nitroureas) (5,34). Lethal toxicity was enhanced when systemic hyperthermia was applied in combination with cyclophosphamide, methyl-CCNU, and carmustine (5). The success of hyperthermia and chemotherapy combinations depends on the temperature increase in the organs for which the drug is used and its subsequent toxicity, all of which are can be influenced by the accuracy of the heating device and the operator.

MODES OF CONVENTIONAL HYPERTHERMIA APPLICATION

Externally Applied Techniques

In the past, single planar transducers were used to apply local heat. A disk shaped piezoelectric transducer (range from 0.3–6.0 MHz in frequency and up to 16 cm in diameter) is mounted above a chamber of cooled degassed water. This device has a coupling chamber which allows water to circulate (3) (Fig. 6). It is coupled to the body via a plastic or latex membrane. Unfortunately, these types of devices are unable to achieve homogenous therapeutic thermal levels. The reason is that this system uses an unfocused energy source. When an unfocused energy source is applied to the skin, the intensity and the temperature will be the highest at the contact point and will subsequently lose intensity as it travels deeper into the tissue. However, by cooling the skin, the “hot spot” is shifted to the subcutaneous fatty tissue that is poorly vascularized. Fat is an insulator and as a result much energy is conserved rather than lost. Furthermore, cooling the skin will produce vasoconstriction which conserves even more heat and facilitates the heating of deeper tissues (3) (Figs. 7, 8a and b). However, even with this strategy adequate temperatures could not be reached. The reason for this is that large tumors often consist of three zones, a central necrotic core, an intermediate zone that is normally perfused, and a marginal zone that has a greater number of vessels due to proliferation induced angiogenesis. Due to the abundance of vasculature on the marginal surface, much heat is dissipated to the surrounding tissue. The relatively avascular center will heat to a higher temperature than the marginal or intermediate zone because there is little dissipation of heat, creating hot spots (Fig. 9) (7,54). Thus, it is not possible to therapeutically heat a tumor with a single planar source. This theory is substantiated by a significant number of clinical trials (7,55–61). Most trials reported that patients had some difficulty with dose-limiting pain, extensive central tumor necrosis, blistering, and ulceration (7). Conversely, a focused ultrasound source localizes energy within the tumor volume while sparing the surrounding tissue (Fig. 10). The use of a focused beam allows for homogenous heating and higher intensity which allows the generation of greater temperatures within the target area. Attenuation and beam divergence cause rapid deterioration of intensity beyond the focal zone (3) (Fig. 11 a and b). Focused ultrasound sources overcome some of the limitations of planar heating. Focusing allows for controlling the amount of heat that is delivered to the poorly perfused areas thus limiting hot spots and some of the side effects. For a heating system to be successful clinically on a large scale, it must account for geometric and dimensional variations of target tissue, possess the ability to heat the sites that need it, and avoid side effects and complications as much as possible (3).

Technological advances in hyperthermia devices have paved the way for better therapeutic options. The use of mosaics or separately controlled transducers allowed better spatial and temperature control to target bulky irregularly shaped tumors. The multielement planar

Table 4. Hyperthermia and Radiation Clinical Trials

Reference/ name of trial	Tumor Entity (stage)	Type of Trial	No. of Patients	Type of Hyperthermia	Results of Control Arm (RT only)[a]	Results of Hyperthermia Arm (RT+HT)[a]	Significance of Results (p<0.05)
(26,32) RTOG	Head and neck (superficial measurable tumor)	Prospective randomized multicenter	106	Superficial (915 MHz microwave)	34% CR	34% CR	–
(25)	Head and neck untreated locoregional tumor	Prospective randomized	65	Superficial (27-12 MHz microwave)	32% DR	55% CR	+
					19% DFS at 1.5 years	33% DFS at 1.5 years	+
(24,33)	Head and neck (N3locoregional tumor)	Prospective randomized	41	Superficial (280 MHz microwave)	41% CR	83% CR	+
					24% LRFS	68% LRFS	+
					0% OS at 5 years	53% OS at 5 years	+
(21) ESHO-3	Melanoma (skin metastases or recurrent skin lesions)	Prospective randomized Multicenter	70	Superficial (various techniques)	35% CR	62% CR	+
					28% LRFS at 5 years	46% LRFS at 5 years	+
(23) MRC/ ESHO-5	Breast cancer (local recurrences or inoperable primary lesions)	Randomized multicenter	306	Superficial (various techniques)	41% CR	59% CR	+
					ca. 30% LRFS	ca. 50% LRFS	+
					ca. 40% AS at 2 years	ca. 40% AS at 2 years	–
(31)	Rectal cancer	Prospective randomized multicenter	143	Deep regional HT (various techniques)	15% CR	21% CR	–
					22% OS at 3 years	13% OS at 3 years	–
	Bladder cancer		101		51% CR	73% CR	+
					22% OS at 3 years	28% OS at 3 years	–
	Cervical cancer		114		57% CR	83% CR	+
					27% OS at 3 years	51% OS at 3 years	+
(28)	Various (recurrent or progressive lesions)	Prospective randomized multicenter	174	Interstitial HT (300-2450 MHz microwave or RF)	54% CR	57% CR	–
					34% OS at 2 years	35% OS at 2 years	
(25)	Gioblastoma (postoperative)	Prospective randomized	79	Interstitial HT	15% OS at 2 years	31% OS at 2 years	+
(29)	Stage IIIB uterine cervix	Prospective randomized	40	Deep regional HT	50% CR	80% CR	+
					45% CR at 3 years	79.7% CR at 3 years	
(27)	Superficial tumors	Prospective randomized	122	EM	42.3% CR	66.1% CR	+
(30)	Uterine cervical	Prospective randomized multicenter	110	RF	68.5% CR	73.2% CR	+

[a] AS = actuarial survival; CR = complete remission; DFS = disease free survival; HT = hyperthermia; LRFS = local relapse free survival; OS = overall survival; RF = radio frequency electric currents; RT = radiotherapy. (Published with permission from Ref 20).

ultrasound applicators met these demands and are capable of treating tumors at depths up to 8 cm. The multisector applicator allows for heating to the edge of the aperture and the acoustic beams are nondiverging in the near field, thus allowing large tumor heating with lateral measurements of 15 × 15 cm. Each of these 16 sectors can be varied from 0 to 100% power to uniformly heat across the tumor. If an area of the tumor is too difficult to treat, more energy

Table 5. Hyperthermia and Chemotherapy Clinical Trials

Reference	Tumor Entity	Type of Trial	No. of Patients	Type of Hyperthermia[a]	Type of Chemotherapy[a]	Results[a]
(36)	Oesophagus cancer (preoperative)	Phase II	32	localHT/ Endoluminal MW	CDDP + Bleo + Cyc	8 CR/13 PR (65% RR)
(37)	Oesophagus cancer (preoperative)	Phase III	20	localHT/ Endoradiotherm	CDDP + Bleo	1 CR/5 PR/4 MR (50% RR); FHR (41.2%)
			20	Control	CDDP + Bleo	0CR/5 PR/0 MR (25% RR); FHR (18.8%)
(38)	Stomach cancer	Phase II	33	RHT/thermotron	Mitomycin + 5FU	3 CR + 10 PR (39% RR)
	pancreatic cancer		22	8 MHz	Mitomycin + 5FU	3 CR + 5 PR (36% RR)
(39)	Pancreatic cancer	Phase II	77	RHT 13.5 MHz	Mitomycin + 5FU +/ - immunostimulation	27..3% survival at 1 year
(40,41)	Sarcomas (pretreated with chemotherapy)	Phase II (RHT 86)	38	RHT/BSD 1000 60-110 MHz	VP16 + IFO	6 pCR + 4PR + 4FHR (37% RR)
		Follow-up	65		VP16 + IFO	9pCR + 4PR + 8FHR (32% RR)
(42,43)	High risk soft tissue sarcomas	Phase II (RHT 91)	59	RHT/BSD 2000 80-110 MHz	VP16 + IFO + ADR	ICR/6pCR + 8PR + 13 MR (47%) OS: 46% at 5 years
(44)	High risk soft tissue sarcoma	Phase III (EORTC 62961)	112	RHT/BSD 2000 80-110 MHz (randomized)	VP16 + IFO + ADR	(08/00)
(45)	Soft tissue sarcoma	Phase II	55	ILP with HT	TNF + IFN + L-PAM	10CR/35PR (82% RR)
(46)	Sarcoma/ teratomas (metastatic)	Phase I/II	19	WBH	IFO + CBDCA	6PR (32% RR)
(47)	Sarcoma (metastatic)	Phase II	12	WBH	IFO + CBDCA + VP16	7PR (58% RR)
(48)	Refractory cancers (advanced or metastatic)	Phase I	16	WBH (Aquatherm)	L-PAM (dose-escalation)	ICR/2PR (19% PR)
(49)	Pediatric sarcomas	Phase II	34	RHT/BSD 2000 80-110 MHz	V16 + IFO + CBDCA	12 NED ('best response')/ 7 CR Duration: 7-64 months
(50)	Pediatric nontesticular germ cell tumours	Phase II	10	RHT/BSD 2000 80-110 MHz	CDDP + VP16 + IFO (=PEI)	5CR + 2PR (70% RR) Six patients alive without evidence of tumour (10-33 months)
(51)	Cervical cancer (recurrences)	Phase II	23	RHT/array-system 70 MHz	CDDP (weekly)	2pCR/ICR + 9PR (52% RR)
(52)	Rectal cancer (Dukes C preoperative)	Phase II	27 35	Intraoperative IHP Control	Mitomycin C Mitomycin C	3 LR 13LR
(53)	Metastatic Sarcoma	Phase II	108	whole body Hyperthermia	IFO/CBDCA/VP16	68% success at 1 year

[a]P = intraperitoneal hyperthermic perfusion; WBH = whole body hyperthermia; 5FU = 5-flurouracil; VP16 = etoposide; IFO = ifosfamide; ADR = Adriamycin = Doxorubin; CDDP = Cisplatin; CBDCA = Carboplatin; Bleo = Bleomycin; L-PAM = Melphan; TNF = tumor necrosis factor alpha; IFN = interferon gamma; p = pathohistological; RR = response rate; CR = complete remission; PR = partial remission; MR = minor response; FHR = favorable histological response >75%; LR = local recurrence; NED = no evidence of disease.(Published with permission from Ref. 20).

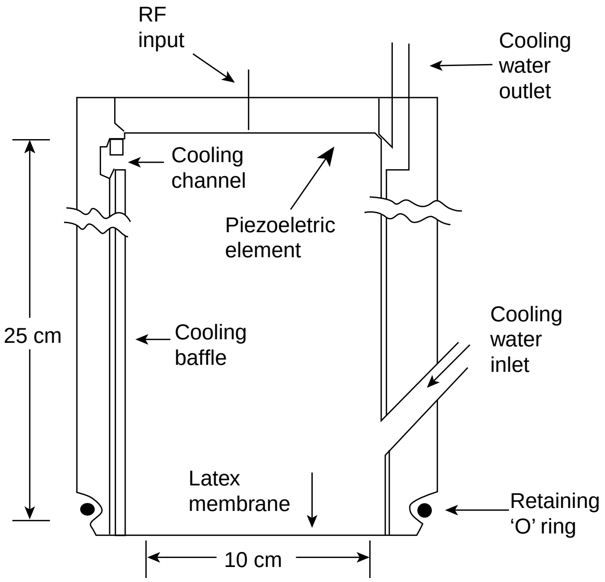


Figure 6. A cross-sectional diagram of a single planar element hyperthermia applicator. The chamber between the latex membrane and the piezoelectric element contains cooled degassed water. During local hyperthermia treatment an ultrasonic conducing gel is applied to the target site that is then coupled to the latex membrane. (Published with permission from Ref. 3).

can be directed to just that target segment. The temperatures can be adjusted in relation to variations in temperature distribution due to blood flow, variations in target tissue morphology, and based on the patient's comfort level. These devices have the ability to contour the energy field to match the tumor outline. These systems generally have two frequencies: 1 MHz (used to heat 3–6 cm) and 3.4 MHz (used for more superficial 2–3 cm) (7,62). Examples of heating temperatures for different ultrasound hyperthermia devices are shown (Table 6). These planar array systems have been adapted to allow for thermoradiation in conjunction with an external beam radiation (7). An extended bolus configuration with an internal reflecting system was created to direct the ultrasound energy into desired tissue. This configuration allows the ultrasound transducer to be outside the radiation beam thus preventing potential interference of the two (7,70).

Another approach to achieving greater spatial control is to use a larger variety of small transducers in a nonplanar geometric configuration. This approach has been used in treating intact breast with cancer (7,71). The patient lies prone while the breast is immersed within the water filled cylindrical applicator (Fig. 12). The cylindrical applicator is composed of eight rings (each ring is 25 cm in diameter by 1.6 cm in height), with up to 48 transducers (1.5 ×1.5 cm plane transducers), which are interspersed around the

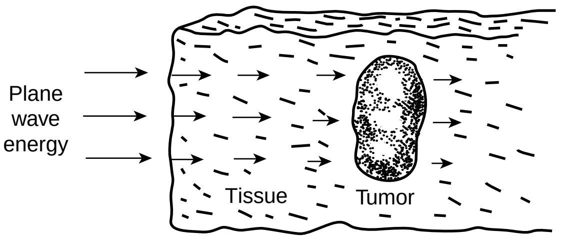


Figure 7. The pattern of ultrasound delivery via plane wave radiation targeting a tumor that is located deep within the tissue. (Published with permission from Ref. 3).

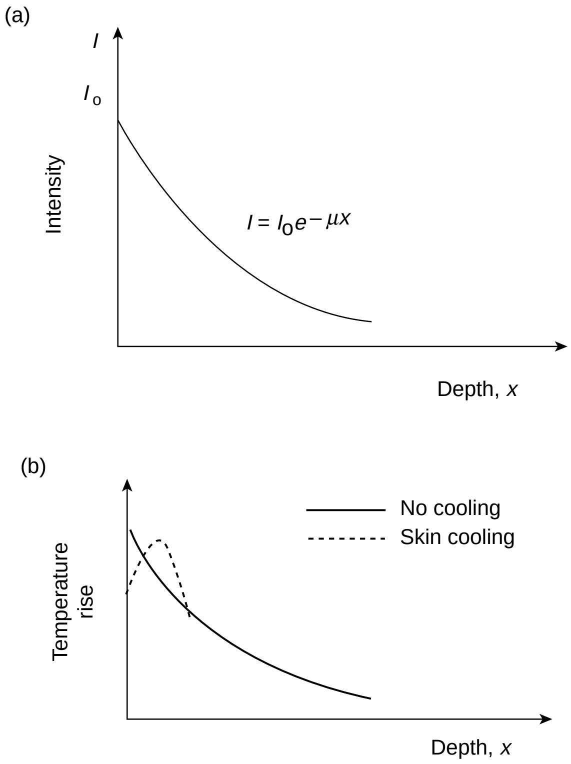


Figure 8. (a) Ultrasound intensity is greatest at the surface. Intensity will deteriorate exponentially due to attenuation as the depth from the surface increases. Published with the permission of (3). (b) Since temperature and intensity are directly proportional, temperature will decrease exponentially as depth increases. Cooling the skin will cause the "hot spot" to shift to the poorly perfused fatty tissue. (Published with permission from Ref. 3).

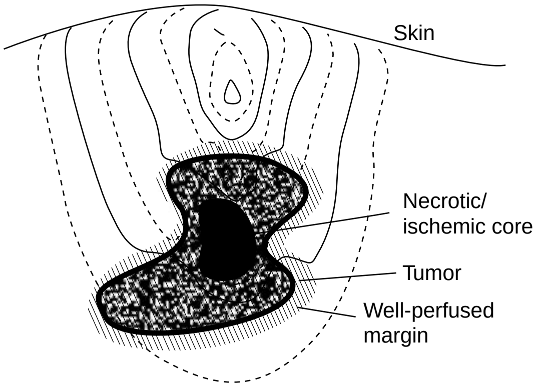


Figure 9. Temperature distribution in a subcutaneous tumor by plane wave transducer. The temperature at the necrotic zone is higher than in the surrounding tissues. (Published with permission from Ref. 3).

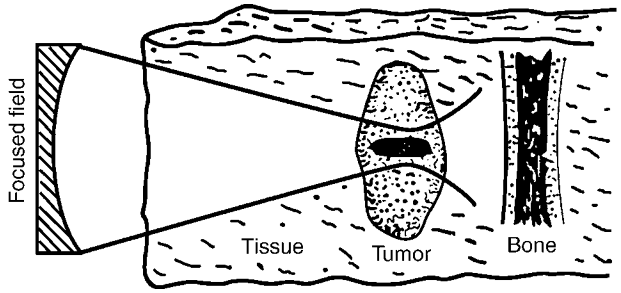


Figure 10. Pattern of radiation with a focused ultrasound beam. The contoured beam localizes the tumor within the focal area while sparing the surrounding tissue (Published with permission from Ref. 3).

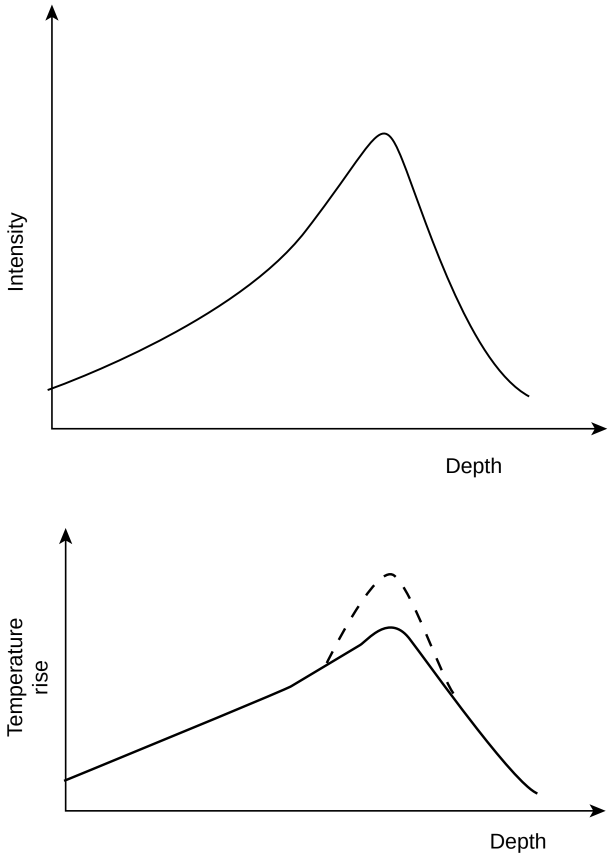


Figure 11. (a) With the use of a focused field, higher intensities can be achieved in target tissue at a greater depth. (Published with permission from Ref. (3)). (b) Since temperature and intensity are directly proportional greater temperatures can also be attained in target tissue at greater depths. (Published with permission from Ref. 3).

ring. The frequency ranges from 1.8 to 2.3 and 4.3–4.8 MHz. The driving frequency and the power can be individually selected within each ring, which allows for better spatial and temperature control. This technique has not yet reached widespread clinical use (7).

The Scanning Ultrasound Reflector Linear Array System (SURLAS), which may soon be implemented in clinical practice allows for 3D power distribution while applying simultaneous external hyperthermia in conjunction with radiation to superficial areas (7,13,72–77). (Fig. 13). The SURLAS applicator consists of two parallel opposed ultrasound linear arrays that aim their sound waves to a V-shaped ultrasound reflector that further organizes and spreads the energy over the scanned target site (7,13). The two arrays operate at different frequencies (1.9 and 4.9). This allows for control of penetration depth through the exploitation of intensity modulation of the two beams (13). The applicator housing this transducer and the temperature regulated water bolus are placed on the patient. This system allows both the radiation and the ultrasonic waves to enter the patient's body concurrently. During the scanning interval, power levels and frequencies in each transducer can be individually regulated, thus allowing for good control over depth penetration and lateral heating (7). This system can treat superficial tumors that are 15 × 15 cm in area and with distal margins up to 3 cm deep (13). However, scan times must be limited to <20 s to avoid transient temperature variations >1 °C (7,73).

Large superficial tumors ranging from 3 to 4 cm deep 20 × 20 cm in surface area have been successfully treated with mechanically scanned planar transducers with 2D motion (7,63) (Fig. 14). This approach can be used in treating tumors in the chest region, which often have a heterogenous thickness and are situated close to bone. Once an ultrasound is launched into tissue, it cannot leave the body; consequently, it will just "bounce" around until it is completely absorbed. If the ultrasound is absorbed by bone or nerves, neuropathies and bone necrosis can occur. Mechanically scanned planar transducer frequencies can range from 1 to 6 MHz. Accurate spatial control has been achieved by controlling the operating frequency and

Table 6. Examples of Clinical Temperature and Response Rates of Certain Hyperthermia Systems[a]

Device	Reference	Number of Patients	Maximum Temperature, °C	Minimum Temperature, °C	Average Temperature, °C	Complete Response Rate, %	Partial Response Rate, %
Scanned ultrasound	(63)	5	45.9	41.1			
	(64)	149				34	36
	(65)	72	44.4	40.0		22	40
	(66)	17	43.1	39.9		24	70
	(67)	15	44	40.4	42.3		
Multielement ultrasound	(68)	147	42.7	38.5	40.4		
Transrectal ultrasound	(69)	14	43.2	40.5	42.2		

[a]Published with permission from Ref. 7.

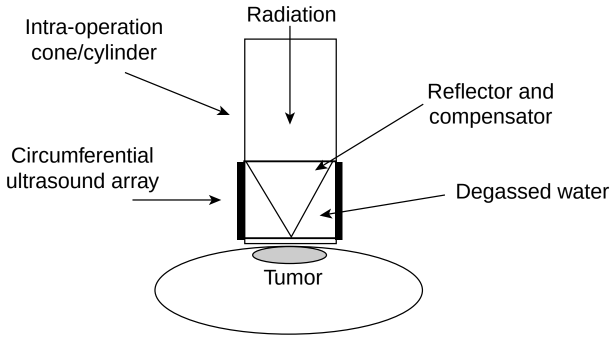


Figure 12. A schematic diagram of an intraoperative multielement transducer system with a circumferential transducer array and reflector configuration. (Published with permission from Ref. 7).

applied power levels, as a function of location, to account for variations of tumor thickness. Separate transducers, which are driven at different frequencies or by time multiplexing the driving frequency of a given transducer between its fundamental and odd harmonic frequencies, are able to create a situation that allows control over penetration depth (7). The penetration depth, as well as the temperature distribution resulting as a function of depth, can be controlled online during the scanning by regulating the frequency amplitude. In the clinical setting, all these biophysical properties must be coupled with the patient's ability to tolerate the treatment to create a functional algorithm (7,63).

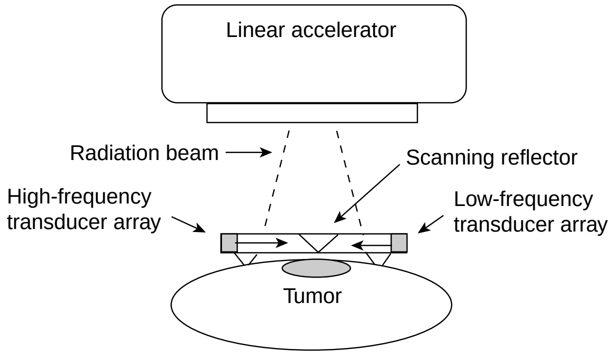


Figure 13. A schematic diagram of a multielement low profile scanning reflector system. (Published with permission from Ref. 7).

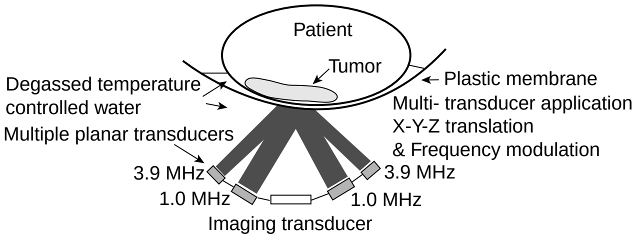


Figure 14. A schematic diagram of mechanically scanned ultrasound system for superficial hyperthermia. (Published with permission from Ref. 7).

Scanned focus ultrasound systems (SFUs) provide the most flexibility for clinical applications (7,64–67,78,79). These systems provide the greatest possibility of overcoming the challenges of tissue heating. The SFUs systems generally use four to six 1 MHz spherically focused transducers each overlapped so that a common focal zone of 3 mm o.d. to treat deep tissue. This focal zone is mechanically scanned in circular or octagonal patterns within the tumor at rates of 20–100 mm·s^{-1}. In order to guarantee that there is a consistency in temperature, scan cycles must be shorter than 10 s. During scanned focused ultrasound hyperthermia treatments, temperature distributions can be controlled by utilizing the measured temperatures to vary the power output as a function of the location. The resolution is determined by a variety of thermometry points, scanning, and computer control speed (7). The regulation of temperature can be controlled by the clinician or the computer (7,80).

External applicator systems for hyperthermia have now been developed that use electrically phased focused transducer arrays. The advantages of using an electrically phased focused transducer array is that it allows for better synthesis of complex beam patterns and the ability to electronically focus and steer. The 3D complex beam-forming techniques result in higher scanning speeds, smaller applicators, and better reliability due to more static parts (7). Examples of electrically phased focused transducer arrays include concentric ring arrays (7,81), sector-vortex phased arrays (7,82), spherical and cylindrical arrays (7,83,84), and tapered phased arrays (7,85).

Intracavitary Techniques

Conventional ultrasonic hyperthermia can be used for intracavitary applications. This modality can be used to treat tumors that are situated deep within the body or with those that situated close to a body cavity. Clinically, prostrate cancer and benign prostate hyperplasia are the best suited for this treatment (7). The transrectal applicator consists of one-half cylindrical transducer segments 10–20 mm o.d. × 10 mm long. It is sectored for better angular control with frequency range of 1.0–1.6 MHz. The transducers are housed in a plastic head; also, a temperature regulated degassed water within an extendable bolus is attached (7,86–88) (Fig. 15). The heating energy is emitted radially from the length of each transducer segment, and the power is applied along the length of the applicator. This technique is able to heat tissues that are 3–4 cm deep from the cavity wall. The temperature controlled water bolus maintains a safe temperature for

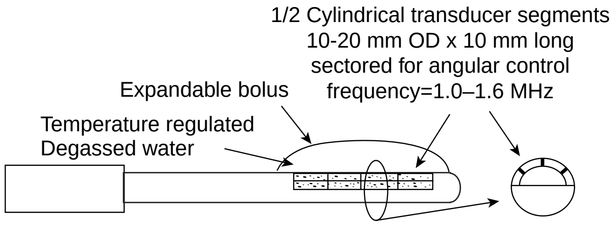


Figure 15. A nonfocused multielement applicator with longitudinal and angular power deposition abilities. This device is used in the treatment of the prostate cancer or BPH. (Published with permission from Ref. 7).

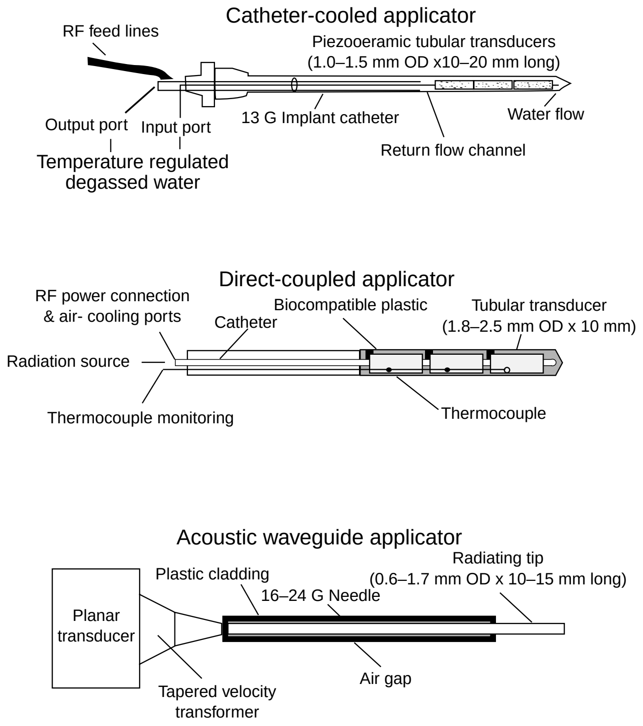


Figure 16. Interstitial hyperthermia catheters. (Published with permission from Ref. 7).

the rectal mucosa. Improved versions of this applicator have added four sectors on each tubular section for 16 channels total. These devices are being fabricated to be compatible with MRI guide protocols (7,89).

Interstitial Techniques

Interstitial techniques are used for treating deep-seated and/or large tumors that are not amenable for surgical resection. Heating sources are implanted into the tumor, thus focusing the energy directly to the site. Interstitial radiation is a standard practice in the treatment of tumors, therefore incorporating adjuvant heat is a logical progression to maximizing treatment. There are three basic designs of interstitial ultrasonic applicators: catheter cooled and direct coupled that consists of tubular piezoceramic transducers, and acoustic waveguide antennas (7) (Figs.16a–c).

Multielement ultrasound applicators with catheter cooling use piezoceramic tubular transducers (1.0–1.5 mm o.d. × 10–20 mm long, with frequency ranging from 7 to 10 MHz) have circulating coolant channels incorporated within the support structures to allow the applicator to be sealed in place within closed end implant catheters (13–14 gauge) (7) (Fig. 16a). These catheters are able to improve control of radial penetration of heat. In addition, it has the ability to control longitudinal power deposition along the length of the applicator (7,90–94). The power to each tubular transducer can be adjusted to control tissue temperature along the length of the catheter. The length and the number of transducers can be selected depending on the desired temperature and longitudinal resolution. This feature is very valuable in that it allows adjustability to tumor geometry variations, blood perfusion variations, and the variation within the tumor tissue. Another advantage of this device is that, unlike microwaves and RF hyperthermia, the power deposition pattern is not limited by the length of insertion or whether other catheters are within the implant. These catheters are more challenging than others for the operator to use skillfully because it is complicated to control both the electronics and the water cooling. Also, small transducers are less reliable. However, it is this complexity that allows for great plasticity in therapeutic temperature distributions (7).

Direct coupled applicators are used to deliver thermobrachy therapy via remote after- loading radiation sources (Fig. 16b). Larger applicator size limits these catheters to few clinical treatments. The implant catheter consists of the transducer and an acoustically compatible housing, which is biologically and electrically insulated. The implant catheter usually ranges from 2.2 to 2.5 mm in diameter. The inner lumen is formed from a catheter that is compatible with standard brachytherapy and commercial after loaders. The transducers have sensors that are able to monitor tissue temperature. In order to conserve size, a water cooling mechanism was not included as part of the catheter. This device is less efficient because transducer self-heating increases the wall temperature and thus reduces radial heating. Therefore, the thermal penetration is sensitive to acoustic frequency (7,95,96). Some studies have shown that integrating an air cooling system to this catheter will allow for better heating penetration (7,95).

The acoustic wave-guide antenna has a minimally invasive 16–24 gauge stainless steel needle that is coupled by a conical tapered velocity transformer to a piezoceramic disk transducer (1.3 cm o.d. operating at 1 MHz) (7,99) (Fig. 16c). The length of the radiating tip can be changed by adjusting the length of the plastic sleeve by 1–1.5 cm. The needle diameter size minutely fluctuates due to Raleigh surface waves propagating from the wave-guide generating flexural vibrations of the needle portion. Acoustic patterns that have been measured demonstrate peaks and nodes in adjacent tissue along the radiating aperture. The temperature of the tissue that is radiated matches the temperature of the radiating antennae. The disadvantages of this system are that the power output is potentially limited for larger or more perfused tumors, and it is difficult to control the longitudinal power deposition (7).

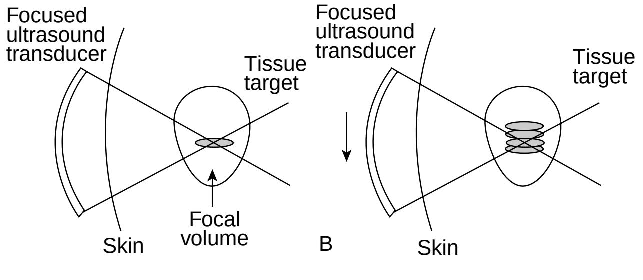


Figure 17. Schematic of HIFU. (a) Illustrates a formation of a single lesion. (b) Illustrates a confluent array of lesions required for a therapeutic effect. (Published with permission from Ref. 98).

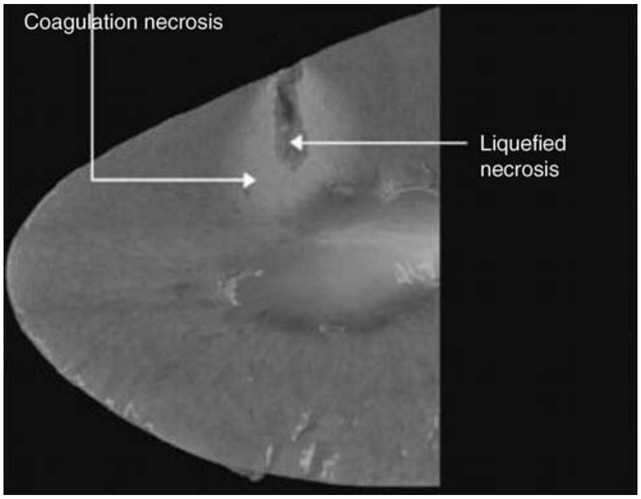


Figure 18. Image of coagulation and liquefied necrosis created with HIFU in an *ex vivo* porcine kidney. (Published with permission from Ref. 108).

A Brief History of HIFU

Using HIFU as an extracorporeal technique of creating coagulative necrosis was first conceptualized in 1942 by Drs. Lynn and Putnam (12,98,99) (Fig. 17a). In 1954, Dr. William Fry was the first to use HIFU to destroy central nervous tissue in the brains of cats and monkeys (12,98,100,101). Later, Frank Fry treated patients with Parkinson's disease and neuromata (12,98,102). Throughout the 1950s and 1960s, HIFU research continued, although it was often plagued with limited success due to lack of technology (103–106). In 1956, Dr. Burov suggested that HIFU can be implemented in the treatment of cancer (12,98,107). Since then, the popularity of HIFU has gradually increased with the advent of better devices and with the success of its use *in vitro* and *in vivo* experimental trials. In current literature, HIFU is categorized as a high temperature hyperthermia because higher temperatures than those used in conventional hyperthermia are required to achieve therapeutic goals.

BASIC PRINCIPLES OF HIFU

The concept of HIFU is similar to that of using a magnifying glass to focus the sun's beams to set fire to some dry leaves. Only the leaves that are in focus will be set on fire, the surrounding ones will be spared (12,98). Likewise, if an ultrasonic beam with sufficient energy is tightly focused, it can be used to elevate temperatures within a target tissue resulting in cell death and coagulative necrosis while sparing the skin and surrounding tissues (98,108) (Fig. 18). Histologically, there is a sharp demarcation between the necrotic tissue that was radiated with HIFU and the healthy surrounding tissue. In the liver, 2 h after exposure, the cells look normal, however, approximately a 10 cell wide rim of glycogen poor cells can be found. After 48 h, the entire area that was radiated will be dead (109).

During HIFU procedures, tissue temperature >56 °C are used because at that temperature irreversible cell death through coagulative necrosis occurs. The main mechanism used is coagulative necrosis via thermal adsorption (110). The other mechanism is cavitation induced damage that is caused by both thermal and mechanical properties of the ultrasound wave (110,111). However, recent studies have been investigating the use of cavitation to enhance the level of ablation and to reduce exposure times. It has been proposed that a focused ultrasound protocol that induces gas bubbles at the focus will enhance the ultrasound absorption and ultimately create larger lesions (110,112). Individual HIFU exposure times can be as little as 1–3 s, while larger volumes may require up to 30–60 s. Individual lesions can be linearly complied to create a clinically relevant lesion (Fig. 17 b). Since individual exposure time is quick, issues (e.g., the cooling effects of blood perfusion) can be considered negligible (7,98,113,114). Therefore, energy transfer and temperature elevation in tissue is considered proportional to acoustic field energy (100). The lesions are cigar-shaped or ellipsoid with the long axis parallel to the ultrasonic beam (12,98). In order to ablate tissue transducer frequency must be between 0.5 and 10 MHz. The higher the frequency, the narrower and shallower the lesion will be. The wavelength ranges from 3 to 0.25 mm. The size of the focal point is determined by the wavelength. Thus, the transverse diameter of the focus is limited to one wavelength and the axial diameter is eight times that wavelength. As a result of this, all generators create a focal size that is 10 × 1 mm. The shape of the lesion is determined by the acoustic properties of the tissue, ultrasound intensity in conjunction with exposure time, and transducer geometry (12). Lesion size is determined by power density at the focus, pulse duration, and the number of pulses. In order to create a well-demarcated lesion the intensity must be $>100\ \mathrm{W{\cdot}cm^{-2}}$, thus being able to reach temperatures that are >65 °C in <5 s (11). Focal peak intensities generally range between 300 and 2000 $\mathrm{W{\cdot}cm^{-2}}$ (7). The ultrasonic waves used in HIFU are generated by piezoelectric elements. In order to achieve high intensity focus ultrasound that is able to ablate tissues three techniques have been found to focus the ultrasound beam: (*1*). spherical arrangement of piezoelements (Fig. 19), (*2*) combination of a plane transducer with an acoustic lens (Fig. 20), (3). cylindrical piezoelements together with a parabolic reflector (11) (Fig. 21).

CURRENT EXTRACORPOREAL DEVICES, INTRACAVITARY DEVICES, AND IMAGING

While there are many devices that are used in experimental trials, few of those are currently used in widespread clinical practice. The two main categories of HIFU devices are extracorporeal and transrectal. Extracorporeal devices have been implemented in experimental trials in many medical fields. Extracorporeal devices use larger transducers, lower frequencies, and longer focal lengths than intracavitary devices (97).

An important factor in clinical application of these devices is the ability to monitor treatment accurately. In current practice, this is accomplished either by using real-time ultrasound (116–118) or MRI (119–122). When

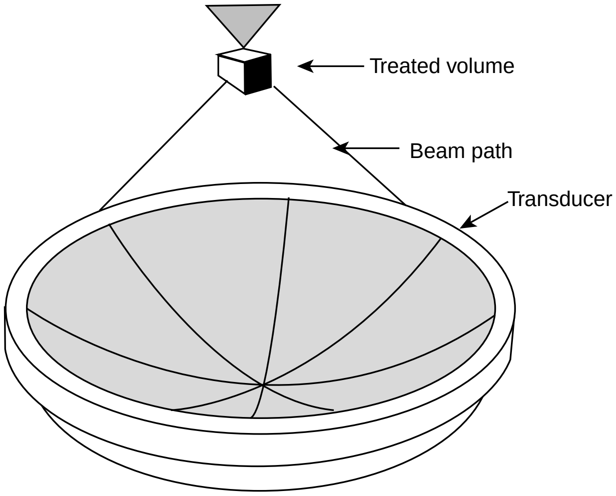


Figure 19. A single spherically curved focused transducer. (Published with permission from Ref. 110).

MR is used to guide HIFU treatments, sublesioning ultrasound exposures are used to identify the target region, local rise in temperatures are used to confirm the position of the ultrasound focus and then higher intensity therapeutic exposures are used for treatment. Currently, several groups are using ultrasound surgery systems that utilize MRI to map temperature elevations online during HIFU procedures (110,120–122). This technique has been used to treat breast tumors and uterine fibroids, and these treatments are in the process of being used clinically in several countries (110,123–125). The MR can effectively use temperature data to determine the parameter of thermal tissue damage (110) and is limited in that it is costly, has lower spin resolution, and because of its technology for producing MR compatible ultrasound equipment required for HIFU is lagging.

When ultrasound is used as a guide, the diagnostic transducer is arranged confocally with the therapeutic transducer and their relationship is fixed. The position of the therapeutic focus can be reliably identified on the diagnostic image. The extent of treatment can be monitored by recording post-treatment gray scale changes on the diagnostic image (98). Ultrasound as a guide is advantageous in that it is less expensive and is more readily accessible, it has faster treatment times, compact sized equipment, and provides a good correlation between observed ultrasound changes and the region of necrosis in the tissue. The disadvantage of using ultrasound as a guide is that image quality is not optimal (98,110,126). Furthermore, ultrasound waves are obstructed by bone and air-filled viscera.

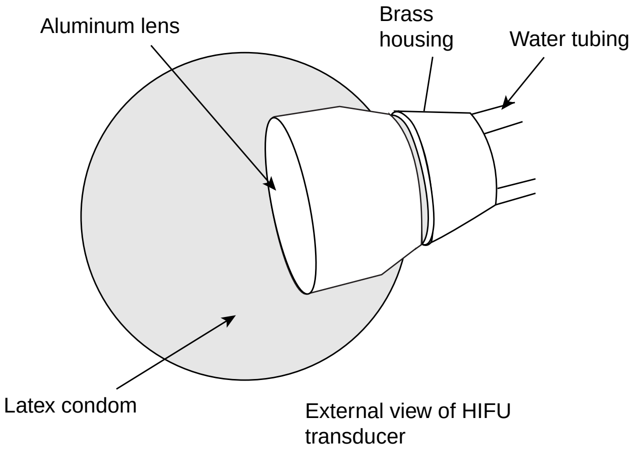


Figure 20. A plane transducer with an acoustic lens used for focusing the ultrasound waves. (Published with permission from Ref. 115).

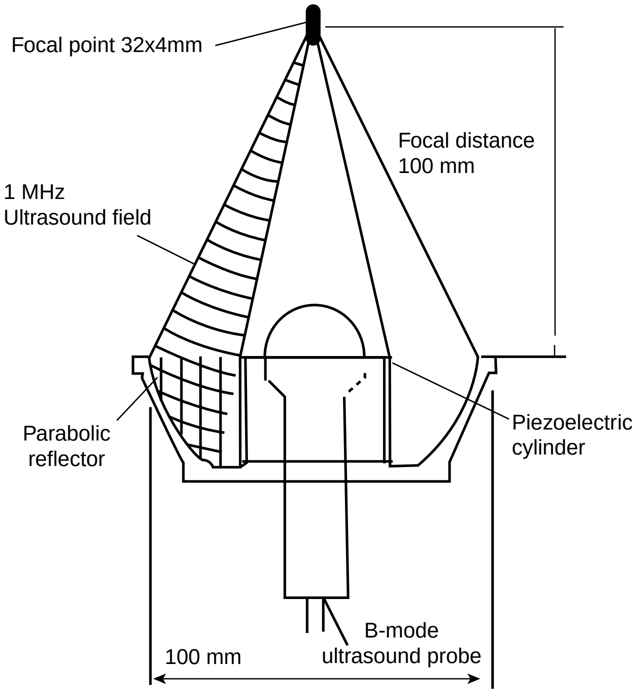


Figure 21. A cylindrical transducer with parabolic reflector. (Published with permission from Ref. 108).

MEDICAL APPLICATIONS OF HIFU

Liver Cancer

While hepatocellular carcinoma is frequently encountered in clinical practice, hepatic metastasis from other primary sources is much more common. Currently, the only definite treatment choice for hepatic metastases is surgery, however, 5 year survival rates are only 25–30%. Arterial embolization is another emerging technique. Therefore, the desire to find a noninvasive technique is preeminent (98). The Chongqing HAIFU device has been used for a couple of years in China to treat a variety of tumors, however, adequate data has not yet been collected (98,127,128) (Fig. 22). The JC-HIFU system (HAIFU Technology Company, Chongquing, PR China) uses an extracorporeal transducer that operates at 0.8–1.6 MHz, the aperture 12–15 cm, focal length 9–15 cm. It operates at Isp of 5–15 kW·cm^{-2}. A diagnostic ultrasound probe (3.5 MHz) is aligned along the same axis as the therapeutic transducer. Both the treatment and diagnostic transducers are placed in a reservoir of degassed water in the center of the

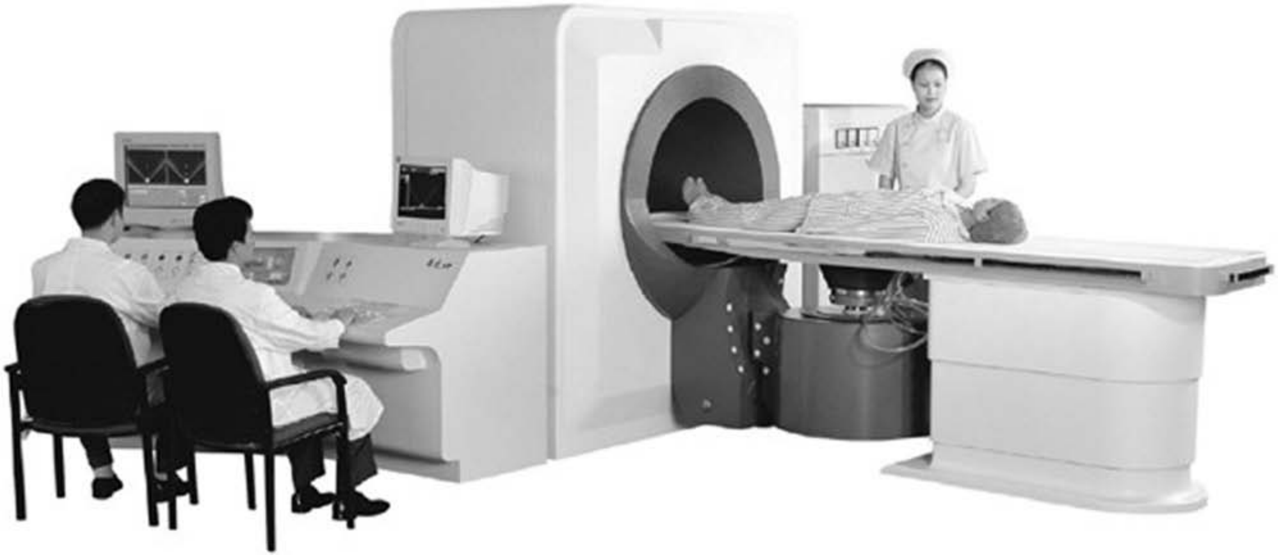


Figure 22. A Hifu System that is used both clinically and experimentally in the treatment of liver metastates. (Published with permission from Ref. 128).

treatment table. The degassed water provides acoustic coupling between the patient and the transducer. Horizontal movement of the transducer is possible along three orthogonal axes of the bed because it is facilitated by the cylindrical gantry at one end of the table. All movement is controlled by the adjacent computer terminals (128). In a recent clinical trial carried out in Churchill Hospital in Oxford, England in conjunction with Chongqing University of Medical Sciences in Chongqing China, 11 patients with liver metastases were treated with the JC-HIFU device. While it is not possible to have a good statistical analysis with such a small subject pool, some general observations were made about the safety of this device. Of the 11 patients treated, 7 out of 11 patients complained of transient pain and 3 out of 11 complained of superficial burns. Out of the 7 patients that experienced pain, oral analgesia brought relief to 6. Burn sites were treated with ice-packs and aloe gel. Two of the three burn sites were only millimeters across. One of the burns was 2 × 3 cm and had healed by the 2 week follow-up period. It appears that from a safety standpoint the JC-HIFU is a feasible treatment option for hepatic metastases, however, larger trials will be needed to determine the true efficacy of the treatment (128).

Another study by Wu et al looked at 55 patients with hepatocellular carcinoma with cirrhosis. Tumor size ranged 4–14 cm in size with an average size of 8.14 cm. Patients were classified according to progression of disease: 15 patients had stage II, 16 had stage IIIA, and 24 had stage IIIC. All patients were treated with an extracorporeal HIFU device similar to the one previously mentioned for the treatment of liver metastases. The average number of treatment applications was 1.69. There were no serious side effects. Imaging following HIFU treatment evaluated for the absence of tumor vascular supply and shrinkage of treated lesions. Serum alpha-fetoprotein returned to normal in 34% of patients. At 6 months, 86.1% of the patients were still alive, at 12 months 61.5% of the patients were still alive, and at 18 months 35.3% of the patients were still alive. The survival rates were the highest in patients who were stage II. Therefore, this study demonstrated that HIFU is a safe option in the treatment of hepatocellular carcinoma (129).

Prostate Cancer

Prostate cancer is one of the common types of cancer in males, and it is frequently the cause of cancer-related death (130). Since physicians are able to detect prostate cancer early, there has been an increase in the number of patients needing treatment. Radical prostatectomy is the treatment of choice in patients who have organ-confined disease and a life expectancy of >10 years. Radical prostectectomy offers excellent results 5 and 10 years after the operation, although there is still risk of morbidity associated with the operation, thus precipitating the need for a noninvasive procedure. Currently, brachytherapy, cryosurgery, 3D conformal radiotherapy, and laparoscopic radical prostatectomy have been implemented with good results (130,131). However, if a cure is not achieved, these treatments cannot be repeated and there is high risk of morbidity associated with salvage radical prostatectomy, thus necessitating the need for another treatment option. In 1995, Madersbacher reported that they were able to destroy the entire tumor within the prostate (98,132). Early reports showed success rates of controlling local tumors at 50% at 8 months and then approaching 90% in later studies (98,133,134). In the later years, as clinicians gained more experience and as technology has improved, treatment of the entire gland was performed (98,135,136).

A recent report was published that looked at 5 year results with transrectal high intensity focused ultrasound in the treatment of localized prostate cancer. One hundred and forty six patients were treated with Ablatherm device (EDAP, Lyon, France). The tablespoon-shaped intracavitary applicator contains both a 7.5 MHz retractable ultrasound scanner for diagnosis and a piezoelectric therapeutic transducer that can be driven at frequencies of 2.25–3.0 MHz. The computer-controlled treatment head is able to move three dimensionally. The applicator can be rotated 45 ° in the rectal ampulla. A cooling device that consists of a balloon containing degassed coupling fluid surrounds the treatment head. Energy can be released from the balloon rectal interface thereby maintaining rectal temperatures <15 °C. Out of 137 patients 6 reported symptomatic UTI, 2 reported chronic pelvic pain, 16 reported infravesicular obstruction, 8 reported grade1 stress incontinence, and 1 reported rectourethral fistula. The success rate of the Ablatherm system is between 56 and 73% (131).

Another study that was published by Uchida et al. performed 28 HIFU treatments on 20 patients to treat localized prostate carcinoma (T1b-2NOMO). A modified Sonoblate 200 HIFU device (Focus Surgery, Indianapolis Ind.) was used in this study. Sonoblate 200 uses a 4 MHz PZT transducer for both imaging and treatment. Each pulse delivery ablates a volume of $2 \times 2 \times 10$ mm^3 in a single beam with 2.5 and 4.5 cm focal length probes. Probes with focal lengths of 3.0, 3.5, 4.0 cm can be used in a split-beam conformation to create lesion sizes of $3 \times 3 \times 10$ mm^3. A cooling device maintains rectal temperatures at <22 °C. In this study, there was a 100% success rate. The UTI-like symptoms were common in the first 2 weeks post-HIFU, but were easily remedied with alpha-blockers and painkillers. One patient had a rectourethral fistula after a second HIFU treatment. Of 10 patients who were still able to attain tumescence prior to the procedure, 3 reported postoperative impotence. It is hypothesized that the reason the Sonoblate 200 is getting superior results to the Ablatherm system is that the treatable focal length is longer in the Sonoblate system. This allows the Sonoblate 200 to treat prostates <50 mL, whereas the Ablatherm can only treat prostates that are <30 mL. However, a controlled prospective study is needed to evaluate the potential reasons for this difference in efficacy (130).

Gynecology

The most common pelvic tumor in women of reproductive age is fibroids. The current surgical options available to manage fibroids are either hysterectomy or myomectomy. Hysterectomy is often not a viable option for women who wish to have children. Myomectomies often result in 50% tumor recurrence in 5 years. Hormone therapy results in temporary reduction in tumor size by 35–65% (115). Therefore, there is a need for a permanent, noninvasive technique to manage fibroids. A device was developed for treating uterine fibroids. The prototypic device aligns a commercial abdominal diagnostic ultrasound transducer with a therapeutic ultrasound intracavitary probe (Fig. 23). This device was constructed to accommodate the specific constraints of the female pelvic anatomy. The transducer contains a 3.5 MHz PZT-8 crystal, 25.4 mm in diameter bonded by an aluminum lens to focus the ultrasound beam. A water-filled latex condom is used for acoustic coupling of the transducer and also has the potential for transducer and tissue cooling. Ergonomics testing in humans demonstrated clear visualization of the HIFU transducer in relation to the uterus, thereby demonstrating a potential for HIFU to treat fibroids from the cervix to the fundus through the width of the uterus. However, this device needs to be tested in treating the uterus in large animal models prior to beginning human trials (115). Extracorporeal devices have been used in a small clinical phase I trials, but the results are still pending (98,137).

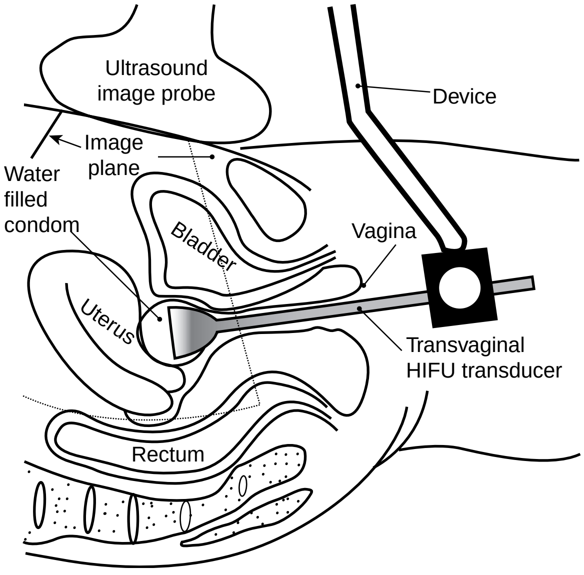


Figure 23. Conceptual diagram of the transvaginal device HIFU in use. The dotted line represents the image plane. Three factors determine what area can be treated by the HIFU transducer: focal length, the range of water stand-off, and the range of mobility once inside the vagina. (Published with permission from Ref. 115).

Breast Cancer

Every year >1 million new cases of breast cancer are diagnosed. Breast cancer is the most common malignancy in women (138,139). In the past, the only options available to women were radical and modified radical mastectomy that included dissection of the axillary lymph nodes. More recently, breast conservation surgery in conjunction with radiotherapy, chemotherapy, and hormone therapy has gained popularity in early stages of breast cancer. However, the change in approach toward a less radical surgery, while being better for a woman's body image, has not really increased long-term survival rates in breast cancer patients (138,140). Other options, such as cryoablation and laser frequency have been studied as minimally invasive approaches, however, these techniques are limited in that they require percutaneous access and that they are only able to treat small masses. In a recent study by Wu et al., (65) women with breast cancer (T1-2, N0-2, M0) verified with biopsy were studied. Patients were divided into a control group that had a modified radical mastectomy, and a group that had extracorporeal HIFU followed by a radical mastectomy. The HIFU system used is the same one described earlier in the treatment of liver malignancies, the JC-HIFU therapeutic system. The therapeutic U.S. beam was produced by a 12 cm in diameter PZT-4 ceramic transducer with a focal length of 90 mm that was driven at a frequency of 1.6 MHz. The ellipsoid focal region dimensions were 3.3 mm along the beam axis, and 1.1 mm along the transverse axis. A real-time imaging U.S. device, the AU3 (Esaote, Genoa, Italy) was used at frequencies of 3.5–5.0 MHz. The diagnostic transducer is placed in the center of the therapeutic transducer. Real-time imaging can accomplish three separate tasks. It can locate the tumor that needs to be treated, it can guide the deposition of U.S. energy into the tumor, and it can provide real-time assessment of the coagulation necrosis during therapy. The results demonstrated that there were no severe side effects. Those that were reported included mild local pain, warmth, and a sensation of heaviness in the affected breast. However, only 4 of the 23 HIFU patients had significant pain to require a 3–5 day course of oral analgesics. Only one patient had a minor skin burn. Pathologic examination of the breast tissue revealed complete coagulative necrosis, and the tumor vasculature was damaged. The immunohistological staining revealed that no expression PCNA, MMP-9, and CD44v6 was found, indicating that the tumor cells had lost their ability to proliferate, invade, and metastasize. Therefore, this study demonstrated the safety and efficacy of HIFU in the treatment of breast cancer (138).

Neurology

Recently, there have been some published studies that propose using large array ultrasound transducers to overcome distortions caused by the skull (110). The goal has been to be able to create an array that can focus to destroy target tissue while preserving surrounding tissue. A 320 element array has been used to focus ultrasound through10 human skulls. This approach is completely noninvasive. This technique is modeled after a layered wave vector-frequency domain-model and uses a hemisphere-shaped transducer to propagate ultrasound through the skull using CT scans as a guide (110,141,142). The ability to focus energy has implications that are not limited to just tumor treatment. It has been shown that focused ultrasound can selectively and consistently open the blood brain barrier (BBB) (110).

Another neurological area that may benefit from HIFU is in the treatment of nerve spasticity and pain. Spasticity, which is signified by uncontrollable muscle contractions, is difficult to treat. In a recent study, HIFU was used to treat and suppress the sciatic nerve complex of rabbits *in vivo*. An image-guided HIFU device including a 3.2 MHz spherically curved therapeutic transducer and an ultrasound diagnostic device were used. A focal intensity of 1480–1850 W/CM2 was used to create a complete conduction block in the 22 nerve complexes. Treatment times averaged 36 s. Gross histological examination revealed blanched nerve complex with lesion dimensions of 2.8 cm^3. Further histological examination revealed the probable cause of nerve block as axonal demyelination and necrosis of Schwann cells. This study illustrates the potential that HIFU may have in the treatment of nerve spasticity (143).

Cardiovascular System

The role of ultrasound in cardiology has been instrumental to the increasing knowledge of the cardiovascular system. Diagnostic ultrasound technology has led to a greater understanding of the anatomy and physiology of the human heart and vascular systems. Over the years, in addition to diagnostic use, the role of ultrasound has been expanded to the therapeutic realm. Both conventional ultrasound and HIFU modalities have been used with varied success in many cardiovascular therapeutic applications. These applications range from harvesting the internal mammary artery for coronary artery bypass surgery to ablation of cardiac arrhythmias. An ultrasonically activated (vibrating up to 55,000 Hz) harmonic scalpel (Ethicon Endosurgery, Cincinnati, OH) produces low heat (<100 °C) thereby effectively coagulating and dividing the tissue and has a wide range of applications in cardiothoracic surgery (144). By using a 1 MHz phased array transducer with an acoustic intensity of 1630 $W{\cdot}cm^{-2}$ or 22547 $W{\cdot}cm^{-2}$ one can successfully create precise defects ranging from 3 to 4 mm in diameter *ex vivo* in porcine valve leaflet, canine pericardium, human newborn atrial septum, and right atrial appendage (145). Cardiac arrhythmia is one area where significant work has been done using ultrasonic hyperthermia for therapeutic purposes. Strickberger et al. demonstrated an extracorporeal HIFU ablation of the atrioventricular junction of beating canine heart after thoracotomy (146). Their experimental system consisted of a polyvinyl membrane covering the heart and lungs. The thoracic cavity was filled with degassed water serving as a coupling medium. A 7.0 MHz diagnostic 2D ultrasound (Diasonics VST Master Series, Diasonics/Vingmed Ultrasound Inc) attached to a spherically focused single piezoelectric element therapeutic ultrasound transducer (1.4 MHz frequency; 1.1 × 8.3 mm focal length and 63.5 cm focal zone) with the maximum intensity of 2.8 $kW{\cdot}cm^{-2}$ was applied during the diastole for 30 s to achieve complete AV nodal junctional block (Fig. 24a–c). Experience with HIFU application is very preliminary and has not been tried for AV nodal ablation in humans yet.

Ultrasound had also been used clinically in the treatment of atrial fibrillation (AF), which is the most common arrhythmia affecting 0.5–2.5% of the population globally. Over the last decade, ablation procedures by isolating the pulmonary veins and eliminating electrical triggers from the atria has become a popular and effective mode of therapy for AF. Traditionally, RF energy has been used as an energy source for ablation. Radio frequency catheter ablation of AF requires good tissue contact, multiple lesions, significant experience and manual skills with long procedure time. Complications related to RF application in AF ablation include pulmonary vein stenosis, atrioesophageal fistula, left atrial rupture due catheter perforation or inappropriate amount of power. The limitations of the existing RF technology could be overcome with the use of HIFU balloon systems (147).

Our group performed the initial work on pulmonary vein isolation in humans using a through-the-balloon circumferential ultrasound (conventional-unfocused) ablation system for treatment of recurrent atrial fibrillation (148). Fifteen patients with drug refractory atrial fibrillation underwent a PVI using a novel transballoon ultrasound ablation catheter (Atronix, Inc) (Fig. 25a–c). The ablation system was composed of a 0.035 in. diameter

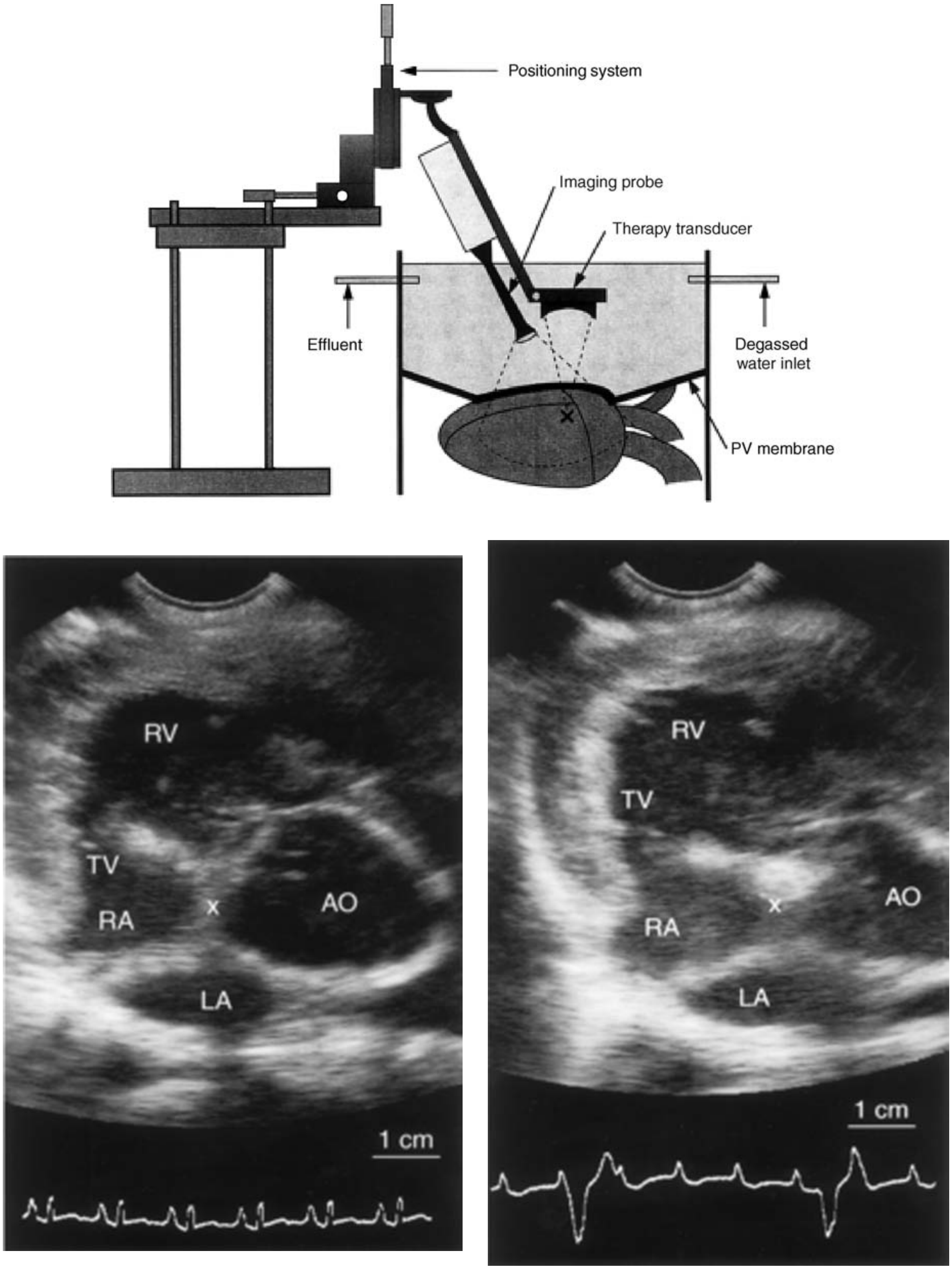


Figure 24. (a) Schematic of the experimental HIFU apparatus. The therapeutic ultrasound transducer is mounted 63.5 mm from the target (X). A polyvinyl chloride membrane covers the heart and the lungs. Degassed water flows in and out of the thoracic cavity at a rate of 600 mL·min^{-1}. Combined diagnostic/ablation transducers are placed into degassed water. (Published with permission from Ref. 146). (b) ECG and Echocardiogram of a canine heart. Prior to ablation of the AV node. (c) After ablation of the AV node using HIFU, the ECG shows complete AV block and the echo image depicts an increased density of the ablated tissue. (Published with permission from Ref. 146).

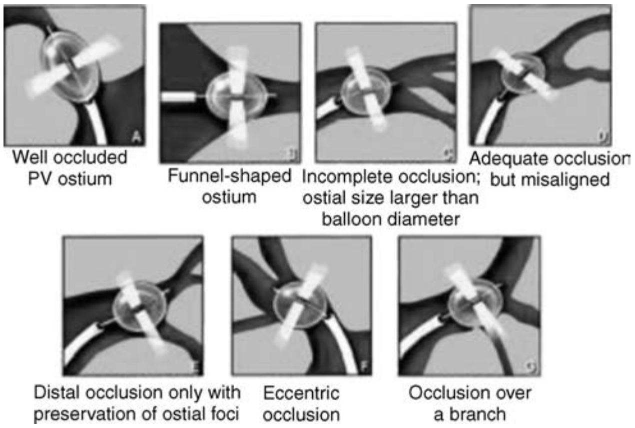


Figure 25. Anatomical pulmonary vein variations and technical limitations. (Published with permission from Refs. 148 and 149).

luminal catheter with a distal balloon (2.2 cm at maximum diameter) housing a centrally located ultrasound transducer (8 MHz). The ultrasound ablation system was advanced over a guide wire (0.035 in., 0.088 cm) to the intended pulmonary vein. The ablation performance and tissue temperature are monitored by thermocouples on the balloon and the therapeutic transducer. The ablation time was 2 min with an additional minute to deflate the balloon. The energy was delivered perpendicularly to the surface of the balloon and ablation at the funnel portion of the pulmonary vein (antrum) could not be achieved with the original design. Additionally, other anatomical characteristics of the target sites like ostial diameter larger than the balloon size, inability to reach the right inferior or other pulmonary vein ostia, ostial instability, early branching of the vein, and eccentric position of the ultrasound transducer in the vein made it difficult to deliver energy effectively (Fig. 26). These technical limitations have been addressed in some of the newer balloon systems where the energy delivery could be accomplished in a divergent angle enabling ablation around the antrum with the tip of the balloon sitting at the pulmonary vein ostium. Early animal studies on HIFU mediated AF ablation have shown promising results with an experimental device that focuses ultrasonic energy via a parabolic balloon, using gas or fluid as a reflector (ProRhythm INC.) (Fig. 25c–e). (149,150).

Since 2003 ~60 patients were treated using this system. With improved catheter design the success rate of AF ablation has increased from ~50 up to 80% without evidence of pulmonary vein stenosis. These preliminary human study results need to be confirmed in larger series. Since there are no large clinical trials on AF ablation with this technology, it is still somewhat premature to predict if HIFU is the complete answer. This device is expected to be released in Europe in 2005 (149,150).

Attempts have been made to harness HIFU for transmyocardial revascularization (TMR) to improve blood supply to damaged myocardium caused by advanced heart disease. Using a 10 cm diameter transducer operating a frequency of 2.52 MHz, intensity of 2300 $W \cdot cm^{-2}$ and pulse repetition period of 40 ms at 50% duty cycle, small channels were successfully created in canine myocardium (151). This shows the potential for future application for HIFU in TMR in a noninvasive fashion.

Other Applications of HIFU

Several branches of medicine have already begun to benefit from the use of HIFU with the prospect of many more applications in the future. Thus far the greatest influence of HIFU has been in oncology, with other fields now exploring and experimenting with HIFU to determine its potential utility. The HIFU has been proposed as a tool for synovectomy in the treatment of rheumatoid arthritis (RA) (98,152) and has been used to control opiate refractory pain in pancreatic cancer patients (98,127) and internal bleeding in organs and vessels (98,153). A hand-held HIFU device has been successfully used to perform vasectomies in dogs as a 1–2 min procedure (98,154).

Future Perspectives in Conventional Hyperthermia and HIFU Use

Heat as therapeutic entity has had a rich history punctuated with many successes and failures. The evolution and integration of therapeutic hyperthermia in the clinical setting have been the product of clinical trials, development of new devices, and education of the medical personnel. Conventional hyperthermia has been used for a long time with many energy sources such as electromagnetic, ultrasound, and microwaves. Similarly, it has been >50 years since HIFU was first conceptualized and actualized. Many subspecialties of medicine have benefitted from the use of hyperthermia. Some of the limitations that were miring conventional hyperthermia and HIFU have only recently begun to be overcome and now these therapies can reach a wider patient population. Both of these techniques share similar obstacles due to the limitations that are inherent to ultrasound. Indeed, ultrasound hyperthermia procedures are limited in that ultrasound cannot propagate through air-filled cavities (e.g., lung or bowel). Consequently, lung tumors other than those that are at the periphery are not likely to be amenable to treatment with HIFU or conventional hyperthermia. Also, tumors that are in close proximity to the bowel or within the bowel wall pose an increased chance of visceral perforation with HIFU use. In addition, other side effects such as pain, soft tissue and bone damage, and skin burns have been reported. Often these side effects can be minimized by varying scan paths, altering frequency, power deposition, or the applicator position (7). The success of both hyperthermia techniques is determined by whether ultrasound energy can be properly directed to the site of interest, or if therapeutic temperatures are achieved, and if other factors can be compensated for, such as hemodynamic changes. Other challenges facing ultrasound hyperthermia systems include the ability to gain control over spatial distribution of heat, tissue temperature monitoring, and improved diagnostic visualizations in order to better treat the tumor site. The HIFU treatment times also need to be shortened. Despite this, treatment times of 1 h for a 2 cm superficial tumor using HIFU is preferable to surgical resection; conversely, at present the same tumor can be treated much

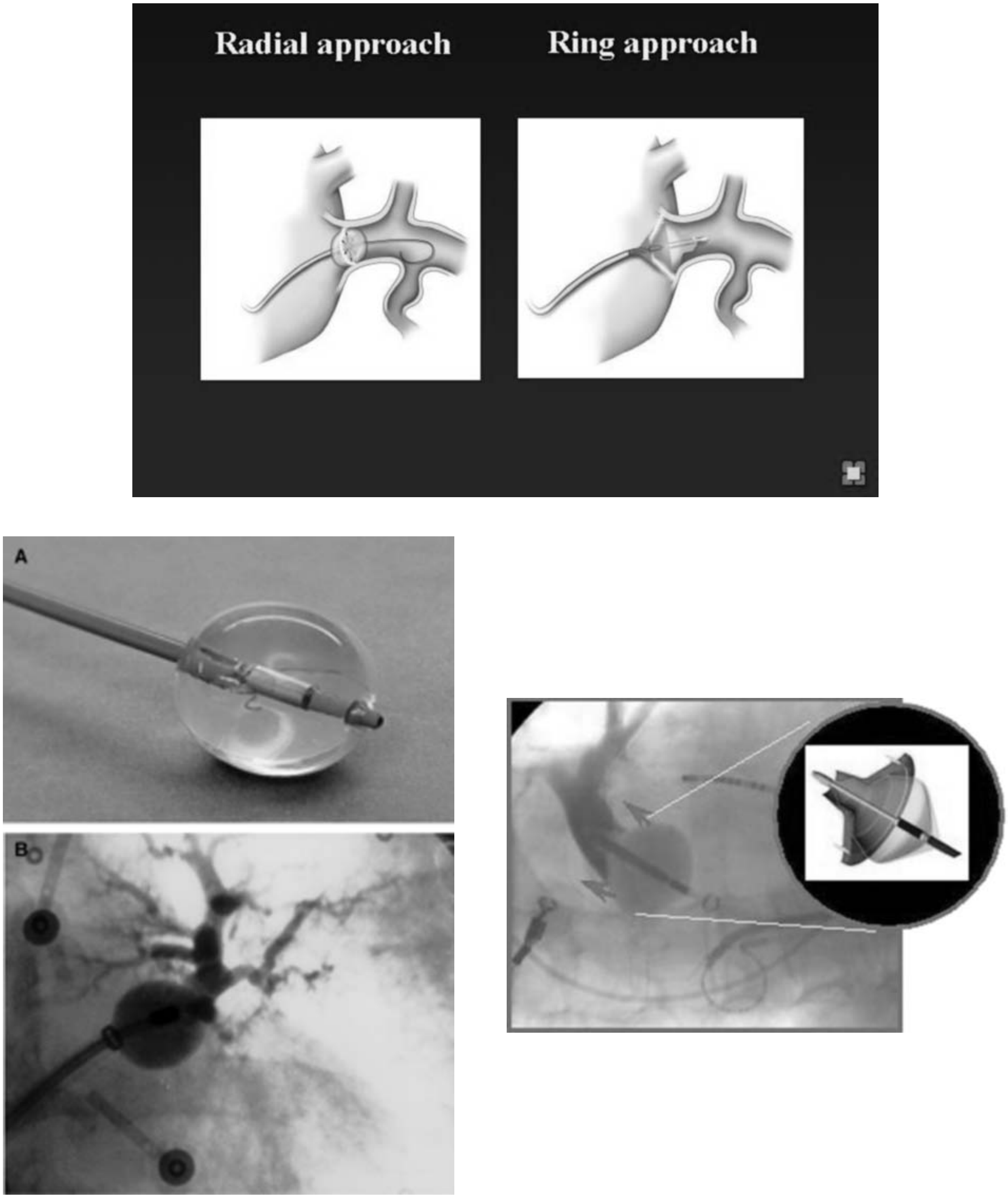


Figure 26. (a) A schematic of the first ultrasound balloon system illustrating the radial delivery of ultrasound energy that is transferred perpendicularly to the tissue. (b) 8F transballoon ultrasound ablation catheter with a central lumen that can accommodate a 0.035 in. guidewire. The distal end of the catheter lodges a cylindrical transducer axially with a saline-filled balloon inflated over it. (c) The balloon is advanced over a guidewire at the ostium of the left upper pulmonary vein. An occlusive pulmonary venogram is done to confirm that the transducer is located at the proximal portion of the vein. Published with permission from Refs. 148 and 149. (d) A schematic of the HIFU device illustrating a ring pattern of ultrasound transmission. The parabolic balloon focuses the ultrasound beam forward. The HIFU catheter is able to create lesions in the antrum away from the lumen of the pulmonary veins thus reducing the likelihood of pulmonary vein stenosis. (e) The HIFU catheter at the ostium of the pulmonary vein.

more quickly with radiofrequency ablation. However, in large tumors longer treatment times are justified because there are very low mortality and morbidity rates associated with the use of HIFU (98).

The benefits of both conventional hyperthermia and HIFU are limitless, but both need more testing in larger patient populations to fully delineate their clinical role. At present, the scarcity of available equipment, the required technical expertise and the lack of thoroughly trained providers, the increased complexity and difficulty involved in using conventional hyperthermia and HIFU when compared to other systems (e.g., electromagnetic) have limited the widespread use of this treatment modality (7). Issues, such as accurate delivery of focused ultrasound to the target tissue, the ability to monitor and control temperatures, and reasonable treatment times, have all been improved upon and will continue to be refined with further testing and research. Imaging modalities like MRI and high resolution CT scan will enhance the precision of HIFU in various system applications. In the future, HIFU could emerge as a good alternative to traditional surgery, and conventional hyperthermia may become a mainstay as an adjunct to chemotherapy and radiation to ultimately improve the arsenal of methodologies available to better treat patients.

BIBLIOGRAPHY

1. Seegenschmiedt MH, Vernon CC. A historical perspective on hyperthermia in oncology. In: Seegenschmiedt MH, Fessenden P, Vernon CC, editors. Thermoradiotherapy and Thermochemotherapy Vol 1. Berlin: Springer Verlag; 1995. pp 3–44.
2. Modern cancer treatment by oncothermia and hyperthermia. 2003. Available at http://www.hot-oncotherm.com/historical.htm. Accessed 2005, Feb 4.
3. Lele P. Hyperthermia Ultrasonics. Encyclopedia of Medical Devices. New York: John Wiley, Sons; 1998.
4. History of Hyperthermia. Available at http://www.starfarm.it/hyperthermia/history.htm. Accessed 2005, April 4.
5. van der Zee J. Heating the patient: A promising approach? Ann Oncol 2002;13:1173–1184.
6. Fesseden P, Hand JW. Hyperthermia therapy physics. In: Smith AR, editor. Medical radiology: Radiation therapy physics. Berlin: Springer Verlag; 1995. pp 315–363.
7. Diederich CJ, Hynynen K. Ultrasound technology for hyperthermia. Ultrasound Med Biol 1999;25:871–887.
8. Bushberg J, Seibert J, Leidholdt E, Boone J. The Essential Physics of Medical Imaging. Philadelphia Lippincott, Williams, and Wilkins; 2002. pp 469–553.
9. Chapelon JY, et al. New Piezoelectric transducers for therapeutic ultrasound. Ultrasound Med. Biol. 2000;26:153–159.
10. Moros E, et al. Thermal contribution of compact bone to intervening tissue-like media exposed to planar ultrasound. Phys Med Biol 2004; 869–886.
11. Madersbacher S, Marberger M. High-energy shockwaves and extracorporeal high-intensity focused ultrasound. J Endourol 2003;17:667–672.
12. ter Haar G. High intensity focused ultrasound for the treatment of tumors. Echocardiography 2001;18:317–321.
13. Novak P, et al. SURLAS: A new clinical grade ultrasound system for sequential or concomitant thermoradiotherapy of superficial tumors: Applicator description.
14. Reinhold H, Endrich B. Tumour microcirculation as a target for hyperthermia. Int J Hypertherma 1986;2:111–137.
15. Vaupel PW, Kelleher DK. Metabolic status and reaction to heat of normal and tumor tissue. In: Seegenschmiedt MH, Fessenden P, Vernon CC. editors. Thermoradiation and Thermochemotherapy, Vol. 1. Berlin: Springer Verlag; 1995. pp 157–176.
16. Raaphorst GP. Fundamental aspects of hyperthermic biology. In: Field SB, Hand JW, editors. An Introduction to the Practical Aspects of Clinical Hyperthermia. London: Taylor and Francis; 1990. pp 10–54.
17. Fajardo LF, Pathological effects of Hyperthermia in normal tissue. Cancer Res 1984;44:4826s–4835s.
18. Sminia P, et al. Effects of hyperthermia on the central nervous system: A review. Int J Hyperthermia 1994;10:1–130.
19. Wondergem J, et al. Effects of local hyperthermia on the motor function of the rat sciatic nerve. Int J Rad Bio 1988;53:429–439.
20. Falk MH, Issels RD. Hyperthermia in oncology. Int J Hyperthermia 2001;17:1–18.
21. Overgaard J, et al. Randomised trial of hyperthermia as adjuvant to radiotherapy for recurrent or metastatic malignant melanoma. The Lancet 1995;345:540–543.
22. Sneed PK, et al. Survival benefit of hyperthermia in a prospective randomized trial of brachytherapy boost (hyperthermia for glioblastoma multiforme). Int J Rad Oncol Biol Phys 1998;40:287–295.
23. Vernon CC, et al. Radiotherapy with or without hyperthermia in the treatment of superficial localized breast cancer: Results from five randomized controlled trials. Int J Rad Oncol Biol Phys 1996;35:713–744.
24. Valdagni R, Amichetti M, Pani G. Radical Radiation alone versus radical radiation plus microwave hyperthermia for N3 (TNM-UICC) neck nodes: A prospective randomized clinical trial. Int J Rad Oncol Biol Phys 1988;15:13–24.
25. Datta N, et al. Head and Neck cancers: Results of thermoradiotherapy versus radiotherapy. Int J Hyperthermia 1990;6:479–486.
26. Perez et al. Randomised phase III study comparing irradiation and hyperthermia with irradiation alone in superficial measurable tumors. Am J Clin Oncol 1991;14:133–141.
27. Jones E, et al. Randomized Trial of Hyperthermia and radiation for superficial tumors. J Clin Oncol 2005;23(13):3079–3085.
28. Emami B, et al. Phase III study of interstitial thermoradiotherapy compared with interstitial radiotherapy alone in the treatment of recurrent or persistent human tumors: A prospectively controlled randomized study by the radiation therapy oncology group. Int J Rad Oncol Biol Phys 1996;34:1097–1104.
29. Harima Y, et al. A randomized clinical trial of radiation therapy versus thermoradiotherapy in stage IIIB cervical carcinoma. Int J Hyperthermia 2001;17(2):97–105.
30. Vasanthan A, et al. Regional hyperthermia combined with radiotherapy for uterine cervical cancers: A multi-institutional prospective randomized trial of international atomic energy agency. Int J Rad Oncol Biol Phys 2005;61(1):145–153.
31. Van der Zee J, et al. Comparison of radiotherapy alone with radiotherapy plus hyperthermia in locally advanced pelvic tumors: a prospective, randomized, multicentre trial. Dutch Deep Heat Hyperthermic Group. Lancet 2000; 1119–1125.
32. Perez CA, Gillespie B, et al. Quality assurance problems in clinical hyperthermia and their impact on therapeutic outcome: a report by the radiation oncology group. International Journal of Radiation Oncology, Biology, and Physics, 16:551–558.
33. Voldagni R, Amichetti M. 1994 Report of long-term follow-up in a randomized trial comparing radiation therapy plus hyperthermia to metastatic lymphnodes in stage IV head and neck patients. International Journal of Radiation Oncology, Biology, and Physics, 28:163–169.

34. Dahl O. Interaction of heat and drugs in vitro and in vivo. In: Seegenschmiedt MH, Fessenden P, Vernon CC, editors. Thermoradiotherapy and Thermochemotherapy, Vol 1. Berlin: Springer Verlag; 1995. pp 103–121.
35. Westermann AM, Grosen EA, et al. A pilot study of whole body hyperthermia and carboplatin in platinum-resistant ovarian cancer. Eur J Cancer 2001;37:1111–1117.
36. Li DJ, Hou BS. A preliminary report on the treatment of esophageal cancer by intraluminal microwave hyperthermia and chemotherapy. Cancer Treatment Reports 1987;71: 1013–1019.
37. Sugemachi K, et al. Chemotherapy combined with or without hyperthermia for patients with oesophageal carcinoma: a prospective randomized trial. International Journal of Hyperthermia 10:485–493.
38. Kakehi M, et al. Multi-institutional clinical studies on hyperthermia combined with radiotherapy or chemotherapy in advanced cancer of deep-seated organs. International Journal of Hyperthermia 6:719–740.
39. Falk RE, et al. Combination therapy for resectable and unresectable adenocarcinoma of the pancres. Cancer 57: 685–688.
40. Issels RD, et al. Ifosfamide plus etoposide combined with regional hyperthermia in patients with locally advanced sarcomas: a phase II study. Journal of Clinical Oncology 1990;8:1818–1829.
41. Issels RD, et al. Improvement of local control by regional hyperthermia combined with systemic chemotherapy(ifosfamide plus etoposide) in advanced sarcomas: updated report on 65 patients. Journal of Cancer Research and Clinical Oncology 1991;117:141–147.
42. Issels RD, et al. Preoperative systemic etoposide/ifosfamide/doxorubicin chemotherapy combined with regional hyperthermia in high-risk sarcoma: a pilot study. Cancer Chemotherapy and Pharmacology 1993;31(suppl. 2) S233–S237.
43. Issels RD, et al. Neoadjuvant chemotherapy combined with regional hyperthermia (RHT) followed by surgery and radiation in primary and recurrent high-risk soft-tissue sarcomas (HR-STS) of adults (updated report). Journal of Cancer Research and Clinical Oncology 1998;124(suppl.) R105.
44. Issels RD. Soft Tissues Sarcomas-What is Currently Being Done. European Journal of Surgical Oncology 21(suppl) 471–474.
45. Eggermont A MM, et al. Isolated Limb Perfusion with high-Dose tumor necrosis factor alpha in combination with interferon gamma and melphalan for nonresectable extremity soft tissue sarcomas: a multicenter trial. Journal of Clinical Oncology 14:2653–2665.
46. Wiedemann GJ, et al. Ifosfamide and carboplatin combined with 41.8 °C whole body hyperthermia in patients with refractory sarcoma and malignant teratoma. Cancer Research 54:5346–5350.
47. Weidemann GJ, et al. Ifosfamide carboplantin, and etopside (ICE) combined with 41.8 °C whole-body hyperthermia in patients with refractory sarcoma. European Journal of Cancer 32A:888-891.
48. Robins HI, et al. Phase I Clinical Trail of Melphalan and 41.8 °C whole-body hyperthermia in cancer patients. Journal of Clinical Oncology 1993;15:154–164.
49. Romanowski R, Schott C. Regionale hyperthermie mit systemischer Chemotherapie bei Kindern und Jugendlichen: Durchfuhrbarkeit und Klinische Verlaufe bei 34 intensiv vorbehandelten Patienten mit prognostisch ungunstigen Tumorerkrankkungen. Klinische Padiatrie 205:249–256.
50. Wessalowski R, Kruck H. Hyperthermia for the treatment of patients with malignant germ cell tumors. A phase I/II study in ten children and adolescents with recurrent or refractory tumors. Cancer 82:793–800.
51. Rietbroek RC, Schilthuis MS. Phase II trial of weekly locoregional hyperthermia and cisplantin in patients with previously irridated recurrent carcinoma of the uterine cervix. Cancer 1997;79:935–942.
52. Takahashi M, Fujimot S. Clinical outcome of intraoperative pelvic hyperthermochemotherapy for patients with Dukes' C rectal cancer. International Journal of Hyperthermia 10:749–754.
53. Westermann AM, Wiedemann GJ. A Systemic Hyperthermia Oncologic Working Group trial. Ifosfamide, carboplatin, and etoposide combined with 41.8 degrees C whole-body hyperthermia for metastatic soft tissue sarcoma. Oncology. 2003;64(4):312–21.
54. Anhalt DP, Hynynen K, Roemer RB. Patterns of changes of tumour temperatures during clinical hyperthermia: Implications for treatment planning, evaluation and control. Int J Hyperthermia 1995;11:425–436.
55. Corry PM, Barlogie B, Tilchen EJ, Armour EP. Ultrasound induced hyperthermia or the treatment of human superficial tumors. Int J Rad Oncol Biol Phys 1982;8:1225–1229.
56. Corry PM, et al. Combined ultrasound and radiation therapy treatment of human superficial tumors. Radiology 1982b;145: 165–169.
57. Harrison GH. Ultrasound hyperthermia applicators: Intensity distributions and quality assurances. Int J Hyperthermia 1990;6:169–174.
58. Marmor JB, Hahn GM. Ultrasound heating in previously irradiated sites. Int J Rad Oncol Biol Phys 1978;4:1029–1032.
59. Marmor JB, Hahn GM. Combined radiation and hyperthermia in superficial human tumors. Cancer 1980;46:1986–1991.
60. Marmor JB, Pounds D, Hahn GM. Clinical studies with ultrasound induced hyperthermia. Natl Cancer Inst Monogram 1982;61:333–337.
61. Marmor JB, Pounds D, Hahn GM. Treatment of superficial human neoplasms by local hyperthermia induced by ultrasound. Cancer 1979;43:188–197.
62. Labthermics Technologies Online.1998–1999. Available at http://www.labthermics.com/hyper.html. Accessed 2005 Jan 26.
63. Anhalt DP, et al. Scanned ultrasound hyperthermia for treating superficial disease. Hyperthermic oncology. vol 2. Proceedings of the 6th international Congress on Hyperthermic Oncology, Tucson, (AZ); 1992. p 191-192.
64. Lele PP. Advanced ultrasonic techniques for local tumor hyperthermia. Radiol Clin N Am 1989;27:559–575.
65. Harari PM, et al. Development of scanned focused ultrasound hyperthermia: clinical response evaluation. Int J Rad Oncol Biol Phys 1991;21:831–840.
66. Hand JW, Vernon CC, Prior MV. Early experience of a commercial scanned focused ultrasound hyperthermia system. Int J Hyperthermia 1992;8:587–607.
67. Guthkelch AN, et al. Treatment of malignant brain tumors with focused ultrasound hyperthermia and radiation: Results of a phase I trial. J Neuro Oncol 1991;10:271–284.
68. Duthkelch et al.
69. Formine et al.
70. Straube WL, et al. An ultrasound system for simultaneous ultrasound hyperthermia and photon beam irradiation. Int J Rad Oncol Biol Phys 1996;36:1189–1200.
71. Lu XQ, et al. Design of an ultrasonic therapy system for breast cancer treatment. Int J Hyperthermia 1996;12:375–399.
72. Moros EG, Fan X, Straube WL. An investigation of penetration depth control using parallel opposed ultrasound hyperthermia. J Accoust Soc Am 1997;101:1734–1741.

73. Moros EG, Fan X, Straube WL, Myerson RJ. Numerical and in vitro evaluation of temperature fluctuations during reflected-scanned planar ultrasound hyperthermia. Int J Hyperthermia 1998;14:367–382.
74. Moros EG, Myerson RJ, Straube WL. Aperture size to therapeutic volume relation for a multi-element ultrasound system: Determination of applicator adequacy for superficial hyperthermia. Med Phys 1993;20:1399–1409.
75. Moros EG, Roemer RB, Hynynen K. Simulations of scanned focused ultrasound hyperthermia. The effects of scanning speed and pattern on the temperature fluctuations at focal depth. IEEE Trans Ultrason Ferroelec Frequency Control 1988;35:552–560.
76. Moros EG, et al. Simultaneous delivery of electronic beam therapy and ultrasound hyperthermia using scanning reflectors: a feasibility study. Int J Rad Oncol Biol Phys 1995;31: 893–904.
77. Moros EG, Straube WL, Myerson RJ. Potential for power deposition conformability using reflected-scanned planar ultrasound. Int J Hyperthermia 1996;12:723–736.
78. Lele PP, Parker KJ. Temperature distributions in tissues during local hyperthermia by stationary or steered beams of unfocused or focused ultrasound. Br J Cancer 1982;45 (Suppl):108–121.
79. Hynynen K, et al. A scanned, focused, multiple transducer ultrasonic system for localized hyperthermia treatments. Int J Hyperthermia 1987;3:21–35.
80. Lin W, Roemer RB, Hynynen K. Theoretical and experimental evaluation of a temperature controller for scanned focused ultrasound hyperthermia. Med Phys 1990;17:615–625.
81. Ibbini MS, Cain CA. The concentric-ring array for ultrasound hyperthermia: combined mechanical and electrical scanning. Int J Hyperthermia 1990;6:401–419.
82. Umemura S, Cain CA. The sector-vortex phased array: acoustic field synthesis for hyperthermia. IEEE Trans Ultraon Ferroelec Frequency Control 1989;36:249–257.
83. Ebbini ES, Cain CA. Experimental evaluation of a prototype cylindrical section ultrasound hyperthermia phased array applicator. IEEE Trans Ultrason Ferroelec Frequency Control 1991a;38:510–520.
84. Ebbini ES, Cain CA. A spherical-section ultrasound phased array applicator for deep localized hyperthermia. IEEE Trans Biomed Eng 1991b;38:634–643.
85. Benkeser PJ, Frizzell LA, Goss SA, Cain CA. Analysis of a multielement ultrasound hyperthermia applicator. IEEE Trans Ultrason Ferroelec Frequency Control 1989;36:319–325.
86. Diederich CJ, Hynynen K. Induction of hyperthermia using an intracavitary ultrasonic applicator. IEEE Ultrason Sympos Proc 1987;2:871–874.
87. Diederich CJ, Hynynen K. Induction of hyperthermia using an intracavitary multielement ultrasonic applicator. IEEE Trans Biomed Eng 1989;36:432–438.
88. Diederich CJ, Hynynen K. The development of intracavitary ultrasonic applicators for hyperthermia: A design and experimental study. Med Phys 1990;17:626–634.
89. Smith NB, Buchanan MT, Hynynen K. Transrectal ultrasound applicator for prostate heating monitored using MRI thermometry. Int J Rada Oncl Biol Phys 1999;33:217–225.
90. Diederich CJ. Ultrasound applicators with integrated catheter-cooling for interstitial hyperthermia: Theory and preliminary experiments. Int J Hyperthermia 1996;12:279–297.
91. Diederich CJ, Hynynen K. Ultrasound technology for interstitial hyperthermia. In: Seegenschmiedt MH, Sauer R, editors. Interstitial and intracavitary thermoradiotherapy. Berlin: Springer-Verlag; 1993. pp 55–61.
92. Hynynen K. The feasibility of interstitial ultrasound hyperthermia. Med Phys 1992;19:979–987.
93. Hynynen K, Davis KL. Small cylindrical ultrasound sources for induction of hyperthermia via body cavities or interstitial implants. Int J Hyperthermia 1993;9:263–274.
94. Lee RJ, Klein LJ, Hynynen K. A multi-element and multi-catheter ultrasound system for interstitial hyperthermia. IEEE Trans Biomed Eng 1999.
95. Deardorff DL, Diederich CJ, Nau WH. Air-cooling of direct-coupled ultrasound for interstitial hyperthermia and thermal coagulation. Med Phys 1998;25:2400–2409.
96. Diederich CJ, et al. Direct coupled interstitial ultrasound applicators for simultaneous thermobrachytherapy: A feasibility study. Int J Hyperthermia 1996;12:401–419.
97. Jarosz BJ. Feasibility of ultrasound hyperthermia with waveguide interstitial applicator. IEEE Trans Biomed Eng 1996;6:1106–1115.
98. Kennedy J, et al. High intensity ultrasound: Surgery of the future? Br J Rad 2003;76:590–599.
99. Lynn JG, Zwemer RL, Chick AJ, Miller AG. A new method for generation and use of focused ultrasound in experimental biology. J Gen Physiol 1942;26:179–193.
100. Fry WJ, et al. Ultrasonic lesions in the mammalian central nervous system. Science 1955;122:517–518.
101. Fry WJ, Mosberg WH, Barnard JW, Fry FJ. Production of focal destructive lesions in the central nervous system with ultrasound. J Neurosurg 1954;11:471–478.
102. Fry FJ. Precision high-intensity focusing ultrasonic machines for surgery. Am J Phys Med 1958;37:152–156.
103. Ballantine HT, Bell E, Manlapaz J. Progress and problems in the neurological application of focused ultrasound. J Neurosurg. 1960;17:858–876.
104. Warwick R, Pond JB. Trackless lesions in nervous tissues produced by HIFU (high-intensity mechanical waves). J Anat 1968;102:387–405.
105. Lele PP. Concurrent detection of the production of ultrasonic lesions. Med Biol Eng 1966;4;451–456.
106. Lele PP. Production of deep focal lesions by focused ultrasound-current status. Ultrasonics 1967;5:105–112.
107. Burov AK. High-intensity ultrasonic vibrations for action on animal and human malignant tumours. Dokl Akad Nauk SSSR 1956;106:239–241.
108. Kohrmann KU, et al. Technical characterization of an ultrasound source for noninvasive thermoablation by high-intensity focused ultrasound. BJU Int 2002;90:248–252.
109. ter Haar G. Robertson D: Tissue destruction with focused ultrasound *in vivo*. Eur Urol 1993;23(Suppl. 1):8–11.
110. Clement GT, Perspectives in clinical uses of high-intensity focused ultrasound. Ultrasonics 2004;42:1087–1093.
111. Holt RG, Roy RA, Edson PA, Yang X. Bubbles and HIFU: the good, the bad, and the ugly. In: Andrew MA, Crum LA, Vaezy S, editors. Proceedings of the 2nd International Symposium on Therapeutic Ultrasound; 2002. pp 120–131.
112. Sokka SD, King R, Hynynen K. MRI-guided gas bubble enhanced ultrasound heating in *in vivo* rabbit thigh. Phys Med Biol 2003;48:223–241.
113. Billard BE, Hynynen K, Roemer RB. Effects of physical parameters on high temperature ultrasound hyperthermia. Ultrasound Med Biol 1990;16:409–420.
114. Kolios MC, Sherar MD, Hunt JW. Blood Flow cooling and Ultrasonic lesion formation. Med Phys 1996;23:1287–1298.
115. Chan A, et al. An image-guided high intensity focused ultrasound device for uterine fibroids treatment. Med Phys 2002;29:2611–2620. Otsuka R, et al. In vitro ablation of cardiac valves using high intensity focused ultrasound. Ultrasound Med Biol 2005;31:109–114.

116. Wu F, et al. Pathological changes in human malignant carinoma treated with high-intensity focused ultrasound. Ultrasound Med Biol 2001;27:1099–1106.
117. Chaussy C, Thuroff S. High-intensity focused ultrasound in prostate cancer: Results after 3 years. Mol Urol 2000;4:179–182.
118. Madersbacher S, et al. Tissue ablation in benign hyperplasia with high-intensity focused ultrasound. Eur Urol 1993;23 (Suppl. 1):39–43.
119. Hynynen K, et al. MR imaging-guiding focused ultrasound surgery of fibroadenomas in the breast: A feasibility study. Radiology 2001;219:176–185.
120. Hynynen K, et al. A clinical noninvasive MRI monitored ultrasound surgery method. RadioGraphics 1996;16:185–195.
121. Hazel JD, Stafford RJ, Price RE. Magnetic resonance imaging-guided focused ultrasound thermal therapy in experimental animal models: Correlation of ablation volumes with pathology in rabbit muscle and VX2 tumors. J Magnet Reson Imag 2002;15:185–194.
122. Weidensteiner C, et al. Real time MR temperature mapping of rabbit liver in vivo during thermal ablation. Mag Reson Imag 2003;50:322–330.
123. Chan AH, et al. An Image-guided high intensity focused ultrasound device for uterine fibroids treatment. Med Phys 2002;29:2611–2620.
124. Tempany CMC, et al. MR imaging-guided focused ultrasound surgery of uterine leiomyomas: A feasibility study. Radiology 2003;226:897–905.
125. Stewart EA, et al. Focused Ultrasound treatment of uterine fibroid tumors: Safety and feasibility of a noninvasive thermoablative technique. Am J Obstet Gynecol 2003;189:48–54.
126. Wu F, et al. Changes in ultrasonic image of tissue damaged by high intensity ultrasound *in vivo*. J acoustic Soc Am 1998;103:2869.
127. Wu F, Wang ZB, Chen WZ, Zou JZ. Extracorporeal High-Intensity Focused Ultrasound for treatment of solid carcinomas: Four-year Chinese clinical experience. Proceedings of the 2nd International Symposium on Therapeutic Ultrasound; July 29-Aug1; Seattle; 2002.
128. Kennedy JE, et al. High- Intensity focused ultrasound for the treatment of liver tumours. Ultrasonics 2004;42:931–935.
129. Wu F, et al. Extracorporeal high intensity focused ultrasound ablation in the treatment of patients with large hepatocellular carcinoma. Ann Surg Oncol 2004;11(12): 1061–1069.
130. Uchida T, et al. Transrectal high-intensity focused ultrasound for treatment of patients with stage T1b-2NOMO localized prostate cancer: A preliminary report. Urology 2002;59:394–399.
131. Blana A, Walter B, Rogenhofer S, and Wieland W. High-Intensity focused ultrasound for the treatment of localized prostate cancer: 5-year experence. Urology 2004;63:297–300.
132. Madersbacher S, et al. Effect of high intensity focused ultrasound on human prostate cancer *in vivo*. Cancer Res 1995;55:3346–3351.
133. Gelet A, et al. Treatment of prostate cancer with transrectal focused ultrasound: Early clinical experience. Eur Urol 1996;29:174–183.
134. Chaussy C, Thuroff S, Lacoste F, Gelet A. HIFU and prostate cancer: The European experience. Proceedings of the 2nd International Symposium on Therapeutic Ultrasound; July 29–Aug 1; Seattle; 2002.
135. Beerlage HP, et al. High-intensity focused ultrasound followed after one to two weeks by radical retropubic prostatectomy: Results of a prospective study. Prostate 1999;39:41–46.
136. Beerlage HP, et al. Transrectal high-intensity focused ultrasound using the Ablatherm device in treatment of localised prostate carcinoma. Urology 1999;54:273–277.
137. Kohrmann KU, et al. High-intensity focused ultrasound for noninvasive tissue ablation in the kidney, prostate, and uterus. J Urol 2000;163 (4Suppl.):156.
138. Wu F, et al. A randomized clinical trial of high-intensity focused ultrasound ablation for the treatment of patients with localised breast cancer. BJC 2003;89:2227–2233.
139. Mc Pherson K, Steel CM, Dixon JM. Breast cancer epidemiology, risk factors, and genetics. Br J Med 2000;321: 624–628.
140. Curran D, et al. Quality of life in early stage breast cancer patients treated with radical mastectomy or breast-converging procedure: Results of EORTC trial 10801. The European Organization for Research and Treatment of Cancer (EORTC), Breast Cancer Cooperative Group (BCCG). Eur J Cancer 1998;34:307–314.
141. Aubry J, et al. Experimental demonstration of noninvasive transskull adaptive focused based on prior computed tomography scans. J Acoust Soc Am 2003;113:84–93.
142. Clement GT, Hynynen K. A noninvasive method for focusing ultrasound through the human skull. Phys Med Biol 2002;47:1219–1236.
143. Foley J, et al. Image-guided HIFU neurolysis of peripheral nerves to treat spasticity and pain. Ultrasound Med Bio 2004;30:1199–1207.
144. Ohtsuka T, et al. Thoracoscopic internal mammary artery harvest for MICABG using the Harmonic Scalpel. Ann Thorac Surg 1997 June; 63 (6Suppl):S10.
145. Lee LA, et al. High intensity focused ultrasound effect on cardiac tissues: Potential for clinical application. Echocardiography 2000 Aug; 17(6 Pt 1):563–566.
146. Strickberger SA, et al. Extracardiac ablation of the canine atrioventricular junction by use of high-intensity focused ultrasound. Circulation 1999;100:203–208.
147. Natale A. Cleveland Clinic Foundation. Personal Interview. March 5, 2005.
148. Natale A, et al. First Human Experience with Pulmonary vein isolation using a through-the-balloon circumferential ultrasound ablation system for recurrent atrial fibrillation. Circulation 2000;102:1879.
149. Saliba W, et al. Circumferential ultrasound ablation for pulmonary vein isolation: Analysis of acute and chronic failures. J of Cardiovasc Electrophysiol 2002;13:957–961.
150. Ayoma H, et al. Circumferential lesion characteristic of high intensity focused ultrasound balloon catheter for pulmonary vein isolation. Heart Rhythm 2004;1:S430.
151. Smith NB, Hynyen K. The feasibility of using focused ultrasound for transmyocardial revascularization Ultrasound Med Biol 1998;24:1045–1054.
152. Foldes K, et al. Magnetic resonance imaging-guided focused ultrasound synovectomy. Scand J Rheumal 1999;28:233–237.
153. Vaezy S, et al. Hemostasis of punctured blood vessels using high-intensity focused ultrasound. Ultrasound Med Biol 1998;24:903–910.
154. Roberts WW, et al. High-Intensity focused ultrasound ablation of the vas deferens in canine model. J Urol 2002; 167:2613–2617.

See also HYPERTHERMIA, INTERSTITIAL; HYPERTHERMIA, SYSTEMIC; THERMOMETRY.

HYPOTHERMIA. See TEMPERATURE MONITORING.

I

IABP. See Intraaortic balloon pump.

IMAGE INTENSIFIERS AND FLUOROSCOPY

Melvin P. Siedband
University of Wisconsin
Fitchburg, Wisconsin

INTRODUCTION

Early fluoroscopic systems used a phosphor-coated sheet or screen to convert incident X-ray photons to light. The radiologist observed the image through a lead glass protective screen. A film camera was often used to record the image. There are two serious disadvantages to this method: the radiologist had to be dark adapted so that image details were hard to see and the collection solid angle of the eye or camera lens was small. The small collection angle meant that the X-ray exposure had to be increased by almost 100 times to have the same diagnostic quality as a conventional radiograph. Photographing the fluoroscopic screen, photofluorography, is no longer used because of the high exposure to the patients. All X-ray images are noise limited by the finite number of X-ray quanta detected and seen. A 50 keV X-ray photon can, at best, produce ~1000 visible light photons, if absorbed by the old-type phosphor. About 5–10% of the incident X-ray photons are stopped and converted to light by the fluorescent screen. The quantum detection efficiency, (QDE) of the screen is the product of the ability to absorb the incident X-ray photon and the probability of emitting light. A thicker screen would absorb more photons, but would also cause more lateral spreading of the light and reduce the resolution of the image. The 5–10% QDE figure is a practical compromise between resolution and sensitivity. If it is assumed that the visible light photons are emitted isotropically, then the lens of the eye or camera subtends only a very small fraction of this light radiation hemisphere. A sheet of film in a radiographic cassette has a phosphor-coated sheet on either side and can collect light photons far more efficiently.

The invention of the image intensifier overcame these objections. The concept of the early image intensifiers was to use a thin, curved, glass meniscus, ~15 cm diameter, coated on the convex side with a scintillator, originally of the same composition as the zinc:cadmium sulfide fluoroscopic phosphor plates, and a photoemitter on the concave surface. Light produced by the scintillator did not have far to travel to excite the photoemitter. This assembly was placed in a vacuum tube and the photoelectrons were accelerated toward an output viewing screen where a small and very bright image was produced. Because the photoemitter was in close optical contact with the scintillator, the collection angle was very large so that a higher radiation exposure was not needed and the image at the viewing screen was bright enough so that dark adaptation was obviated and fine details could be seen. An optical viewer comprising an objective lens and relay lens, similar in design to a submarine periscope, was used to observe the image.

Modern image intensifiers use an epitaxially grown scintillator of CsI with the photoemitter deposited directly on the surface of that layer. Because the epitaxial layer can be made much thicker than the powdery deposit of the older type scintillator for the same resolution, the new image intensifiers require less exposure–image than conventional radiographs. The thicker layer has a higher QDE than the fluoroscopic screens (Fig. 1).

Many modern image intensifiers use a thin curved steel window on which the scintillator and photoemitter are deposited. This geometry eliminates the X-ray scatter produced by the glass window of the older tubes and improves image contrast. Larger tubes up to 35 cm sensor diameter have been made with variable magnification or zoom capability. The window of thin steel is the flash with a coating of aluminum or other metal and etched to create a domain structure, similar to the domains seen in the zinc coating of galvanized steel. The domains here are very small, ~ 100 μm diameter. Cesium iodide is then vapor deposited on this surface and forms epitaxially (i.e., crystal growth follows the orientation of the metallic substrate), and grows as a collection of optically isolated fibers. This scintillator can be made quite thick with little light spreading. The greater thickness means higher quantum efficiency (i.e., a measure of the fraction of incident X-ray photons converted to light photons), when compared to conventional phosphor plate scintillators. A thin layer of silver and antimony is vapor deposited on top of the scintillator and the final sensitization is accomplished by depositing cesium from heated tubes or reservoirs after assembly in the vacuum tube. The AgCs:Sb photoemitter on the surface of the scintillator is similar to the photocathode of a photomultiplier (PMT), tube.

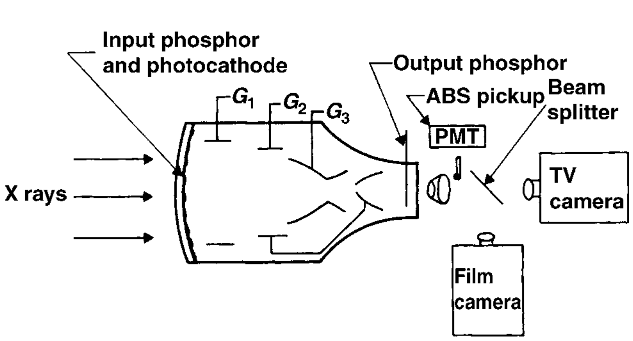


Figure 1. In the image intensifier, X rays strike the input phosphor screen, thus generating light. Light stimulates the photocathode to emit electrons, which are accelerated through 25 kV to strike the output phosphor screen. Brightness gain is due to both geometric gain and electronic gain.

The vacuum tube uses a series of metal cylinders between the photoemitter and the output phosphor. The photoemitter, cylinders, and output phosphor are connected to voltage sources to create shaped electric fields. The field potential at any point affects an electron beam in the same way that the index of refraction of a glass lens affects a beam of light. These cylinders and their potentials form electrostatic lenses to focus the photoelectrons to produce the small and bright image on the output phosphor. Like optical lens elements, the metal cylinders can make compound electrostatic lenses so that focal length can be varied to change the size of the output image: variable zoom.

The brightness gain of an image intensifier is a measure that compares the brightness of the image at the small output screen to that of the older fluoroscopic screens. This gain is a result of the added energy imparted to the photoelectrons, the higher probability of stopping an incident X-ray photon, and the effect of compressing the large input area signal to the small area of the output screen. The older term, brightness gain, has been replaced by conversion efficiency. This term is defined as the output luminance in candela$\cdot$m^{-2} for an input exposure rate of 1 mR$\cdot$s^{-1} or 10 μGy$\cdot$s^{-1}. If assumed it is that a single 50 keV input X-ray photon has a 50% probability of interacting with the input scintillator and producing light (i.e., QDE is 50%) and produces 2000 visible light photons, of which $\sim$ 1000 reach the photoemitter. About 100 of these will produce electrons that will be accelerated by the electric field across the tube and strike the output phosphor as 25 keV electrons. Each 25 keV photoelectron has about a 10% probability of converting its energy into 2.2 eV visible light photons. The QDE, quantum detection efficiency, of the image intensifier is assumed to be 50% because of the thicker CsI scintillator.

An input exposure rate in the diagnostic range produces $\sim$200,000 X-ray photons$\cdot$mm$^{-2}\cdot$s^{-1}, converts to 100,000 light photons/X-ray photon. In the visible range, $\sim$10^{10} photons/mm^2 have a light intensity of 1 candela/m^2. Substituting in the above, a 1 mR (10 μGy) input would produce $\sim$1 candela$\cdot$m^{-2} for an output screen the same size as the input scintillator. This is $>$200 times brighter than the older fluoroscopic screens. Because the output screen area of an image intensifier is 25 mm diameter for an input screen of diameter of 220 mm, the same number of output light photons are emitted from a smaller area resulting in an additional gain factor of $\sim$75. This yields a total brightness gain of $>$15,000 over the old fluoroscopic screen!

A fluoroscopic television system is used for dynamic studies, to enable the viewer to see how a contrast agent is swallowed, how blood flows, to locate objects for a surgical procedure, and so on. During the study, an image record (formerly a film record) could be made for later examination using a video recorder or computer. A rapid succession of images could be recorded and then viewed to find the few images revealing the particular problem, for example, how a heart valve functioned or to diagnose an eroded region in the esophagus as the contrast medium sped by. Or a single image could be made to show the problem for later correction.

A fast optical lens optically collimates the small output image of the image intensifier. Collimation, in this case, means that the optical object is at the focal plane of the collimating lens and its image is focused on an infinite distance away. Any optical device with its own objective lens, such as a TV or film camera, could be aimed through the collimating lens and its image size would be the original intensified image size times the ratio of (objective focal length)/(collimator focal length). The lenses could be separated by several centimeters before image vignetting occurs. A beam splitting mirror can be interposed on the collimator and the objective lenses so that image light is simultaneously apportioned between a TV and film camera. A small mirror or prism and lens could sample the light and form an image over a small hole in the cover of a PMT tube. The hole would permit only light from the center of the image to reach the PMT. The output of the PMT is used to control the X-ray tube current to maintain constant image brightness for continuous viewing or for automatic exposure control of one or a sequence of recorded images. This automatic brightness stabilizer scheme is similar in concept to automatic exposure control of most digital cameras.

The quality of an X-ray image is a compromise between exposure to the patient and the noise or “graininess” of the image. Most images are made using the ALARA principal (As Low As Reasonably Achievable). For any given X-ray exposure, there are a finite number of X-ray photons incident on the patient and then, through the patient, incident on the image sensor. The statistics of photons–area follow the Poisson distribution, so that the variance (noise) is the square root of the average number of photons in a pixel (picture element). To produce a second image having twice the linear resolution (detail) as the first requires four times the exposure. Because the eye averages exposure time $>$0.2 s, to record an equivalent image in 0.02 s requires an exposure rate 10 times greater. X-ray exposure requirements are determined by the by the X-ray absorption of the patient, diagnostic needs and are different for continuous viewing (real-time fluoroscopy), a sequence of images (video or motion pictures), or single images for later diagnoses.

Before the rapid growth and improvement of digital cameras and computer technologies, still and motion picture film cameras were the only practical means to record images. Because of differences of integrating capability of the eye and detail required of the recorded film images, the X-ray beam current, pulse width (exposure time/image), and so on, the ratio of transmission/reflection of the beam splitting mirror, and other operating parameters, must, be adjusted to obtain the required image quality for each image application requirement while minimizing total exposure to the patient.

Most modern systems use charge-coupled image sensors with a high dynamic range and pulse the X-ray beam current to the required level while digitally recording video images. Computer display of selected images or a dynamic sequence of images has largely displaced motion picture film techniques. Video tape recorders are used in simple systems. The automatic brightness control–automatic exposure control of the digital system uses signals from selected image areas derived from the computer image to optimize image quality in the region of interest.

BIBLIOGRAPHY

Further Reading

Gebauer A, Lissner J, Schott O. Roentgen Television. New York: Grune & Stratton; 1966.

Siedband MP. Image storage subtraction techniques and contrast enhancement by electronic means. Symposium on the Physics of Diagnostic Radiology; University of California, June 1968.

Siedband MP. Image intensification and television. In: Taveras, Ferrucci, editors. Radiology, Diagnosis, Imaging, Intervention. Chapt. 10. New York: Lippincott; 1990.

Siedband MP, Duffy PA. Brightness Stabilizer with Improved Image Quality, US patent No. 3,585,391. Accessed 1971.

Siedband MP. X-ray image storage, reproduction and comparison system. US patent No. 3,582,651, 1971.

IMAGING, CELLULAR. See Cellular imaging.

IMAGING DEVICES

Mark J. Rivard
Tufts New England Medical Center
Frank Van den Heuval
Wayne State University

INTRODUCTION

Historically, external radiation treatment of deep-seated malignancies was performed using ortho-voltage equipment. The radiological characteristics of these beams caused maximum dose deposition to occur on the skin of the patient. At that time, skin damage was the limiting factor for dose delivery to the tumor. When the skin turned red due to radiation damage (erythema), the physician had to find another area or portal through which to deliver radiation. The portal was then defined by its orientation and the surface of skin it irradiated.

Nowadays, the treatment is performed with higher photon energies that permit a skin-sparing effect (i.e., the dose at the skin is lower than that deposited a few centimeters deeper) due to the absence of electronic equilibrium. The historic name "portal" still denotes a radiotherapy treatment beam oriented for entry within a patient. Physicians verify whether the treatment is correct using megavoltage treatment beams (4–20 MV photons) as an imaging tool. A transmission image, obtained much like a diagnostic transmission image, provides information describing the patient anatomy and gives clues on the beam orientation and positioning, but also on the extent and shape of the treated area (or portal field). As such, portal imaging is the most direct manner to confirm accuracy of treatment delivery.

Traditional portal verification is done using radiographic films, much like the classical diagnostic films. Films are positioned at the beam exit side of the irradiated patient. Portal image analysis involves comparison with a simulation image that is typically obtained using diagnostic quality X rays (60–120 kV photons). The simulation image serves as the reference image, showing anatomical information clearly and delineating the intended treatment field. Comparison of the simulation image with the portal image is complicated due to the inherent poor quality obtained when imaging using high energy photons (1). The whole procedure of patient positioning, artifact removing, imaging processing, and evaluation using film represents a significant fraction of the total treatment time. This procedure increases the workload per patient, and as a result, the number of images taken is minimized due to economic concerns rather than concerns for efficiency or treatment quality. Indeed, studies have demonstrated that weekly portal image verification, which is the current clinical standard, does not guarantee accurate treatment setup for a population of patients (2).

Portal imaging differs considerably from diagnostic transmission imaging. The main difference is the photon energies used to generate the images. In diagnostic imaging, photons having energies ranging from 50 to 120 kV interact in patients primarily via the photoelectric effect. The cross-section for these interactions is highly dependent on the atomic number of the medium in which they traverse: A higher atomic number increases the probability of interaction. The average atomic number of bony anatomy is higher than that of soft-tissue, yielding good contrast for the bony anatomy. At treatment energies (1–10 MeV) the predominant photon interaction is Compton scattering. The cross-section for this interaction is largely dependent on the media density, and the resulting image will show the largest contrast when large differences in density are present. In practice, this means that differences in soft tissues will contribute most to the visible signal.

These considerations imply that the dynamic range of an electronic portal imaging detector (EPID) is used to obtain information on soft-tissue variations (3), divergence effects (4), scatter contributions (5), field-edge information, and in the case of fluoroscopic imagers: vignetting and glare. With the exception of field-edge information, all of these factors are nonlocalized and tend to change gradually within an image. Not only are these features slowly varying, but they also have a large contrast-to-noise ratio (CNR) compared to the clinically important bone–soft-tissue contrast.

The EPIDs permit the same tasks as film-based imaging, but increase the efficiency and provide added value by using digital imaging techniques. The EPIDs are devices that electronically capture the photon energy fluence transmitted through a patient irradiated during treatment, and allow direct digitization of the image. This image is then immediately available for visualization on a computer screen and electronic storage. When the treatment verification process uses EPIDs, departmental efficiency is increased and quality is improved at the same cost as when using film-based imaging.

Proposals to generate electronic images started in the beginning of the 1980s mainly through the work of Bailey et al. (6), who used systems based on video techniques. This seminal work was then further developed toward more clinically applicable systems by Shalev and co-workers (7), Leong (8), Munro et al. (9), and Visser et al. (10). All of these systems were the basis for the first generation of commercially available EPIDs. They all combined an analog camera with a fluorescent screen generating the optical coupling using a mirror system. Wong et al. (11) replaced the mirror system with optical fibers (one for each pixel). The technology developed further, and is described below in greater detail.

PHYSICAL ASPECTS OF ELECTRONIC PORTAL IMAGING TECHNOLOGY

Camera-Based Detectors

The initial experience using EPIDs was obtained using camera-based systems. Again similar to film-based portal imaging, the camera-based systems measured the photon energy fluence exiting the patient. However, phosphorescent and fluorescent screens replaced the film, and a mirror was oriented at an angle of 45° to reflect the screen toward a video camera. Subsequently, the image was digitized. Because of the intrinsically low-detector efficiency, bulky detector size, and poor image quality, this technology has now become outdated in comparison with more sophisticated technologies.

The low-detector efficiency was due to many limitations in the signal path from screen to computer. Screen conversion efficiency was not ideal when using Gd_2O_2S. In addition, $<0.1\%$ of the light emitted reached the video camera, due to the poor light collection efficiency of the video camera lens. This low rate of signal collection was subsequently impacted by competing electronic noise from the camera in close proximity to the operating linear accelerator (linac). Also, image acquisition typically required a full treatment fraction as compared to the technique using partial fraction irradiation that is commonly used for radiographic portal imaging. Due to the camera-based system detector orientation, rigid positioning of the large mirror was crucial. Changes in linac gantry rotation could cause apparent changes in patient positioning due to physical sag of the camera and mirror mounting system. Furthermore, image quality was suboptimal due to the large lenses required to focus the light signal to the video camera. Degradation of spatial resolution, field uniformity, signal-to-noise (SNR), and field flatness all contributed to minimizing the utility of this detector type.

LIQUID IONIZATION CHAMBERS

The liquid ionization chamber (LIC) is based on a design proposed by Wickman (12), who proposed to use liquid as an ionization medium to increase the efficiency of ionization chambers. Indeed, the introduction of isooctane increased the signal level by over a factor of 10, but also deleteriously increased the recombination of the electrons due to their low mobility. A first prototype was built by Meertens et al. (13) using two printed circuit boards with perpendicular electrode strips. This resulted in a 30×30 matrix of ionization chambers, and was further refined by van Herk et al. (14) to include 128×128 and finally 256×256 matrices.

To obtain an image, the matrix is scanned row by row, by successively switching high voltage to different row electrodes and measuring all column electrodes. The ionization chamber polarizing voltage is typically 300 V, which is comparable to the voltage applied over a regular megavoltage ionization chamber. The typical current produced by the chamber is of the order of 100 pA. Due to the high voltage switching there is a limit on the speed with which the image may be obtained.

In most of the commercially available imagers, ~1 s is required to readout the complete matrix. Figure 1 shows a schematic diagram of a EPID. The LIC is efficient in that it is able to obtain information constantly in between readouts. The low recombination rate ($\alpha \simeq 4.510^{-16}$ m^3/s) of the ions in the liquid makes that the signal accumulates during radiation and provides an averaging effect.

Multiple Detector Combinations

An alternative way to obtain two-dimensional (2D) transmission images is to use a line detector much like the ones found in computer–tomography devices. They consist of a

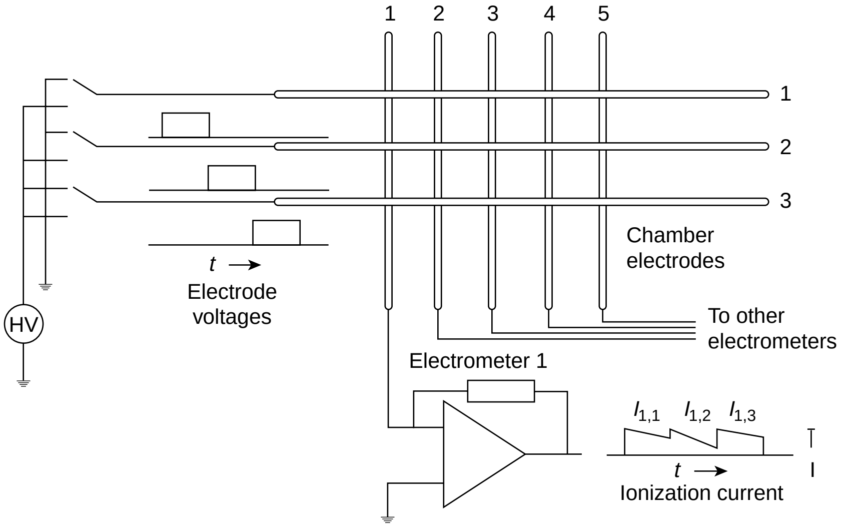


Figure 1. Implanted gold seeds, imaged using a flat-panel portal imager.

line of point detectors, which usually contain a phosphorescent material and an optical light detector (15). Alternatively, Lam et al. (16) constructed a device containing 256 silicon diodes. The line of detectors is scanned through the field in a mechanical fashion. However, this approach is time intensive and not appropriate for clinical techniques such as respiratory gated radiotherapy for treatment of lung cancer where the exiting photon energy fluence is of a dynamic nature.

Flat Panel Technology

The advance of flat-panel displays, where the use of amorphous silicon created surfaces that locally behaved as a crystalline material, allowed for lithography of integrated circuits. The same thin-film technology (TFT) was used to generate photodiode circuits detecting optical light. The TFT is deposited on a glass substrate of ~1 mm thick as is shown in Fig. 2. One of the major advantages of these circuits is that they are highly radiation resistant and can be placed directly in a radiation beam. As with computer integrated circuit chip technology, the TFT EPID can be etched with a resolution of a few micrometers, permitting construction of a large detector matrix. As the photodiodes only detect visible light, a phosphorescent screen is used to perform the conversion much as for camera-based EPIDs. The TFT EPID is nonconducting during the radiation. To read out the TFT, a voltage bias is applied to allow collected charge to flow between the photodiode and an external amplifier. An amplifier records this charge, which is proportional to the light intensity. The TFT EPID array has a maximum readout rate of 25 Hz. In comparison to the camera-based EPID system, the large TFT detectors are designed to be in direct contact with the conversion screen, thus eliminating the poor optical coupling and efficiency intrinsic to the camera-based systems.

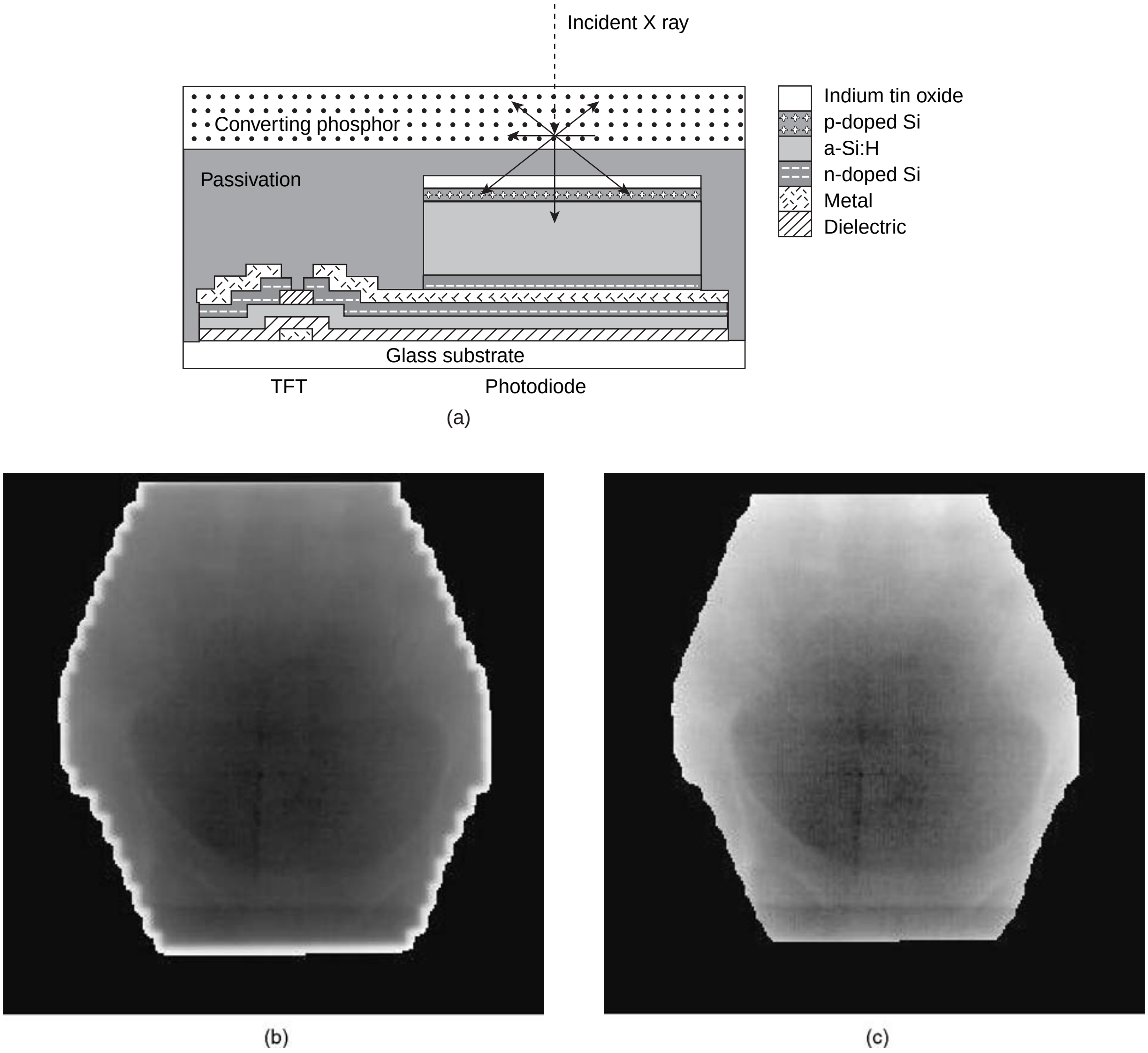


Figure 2. Schematic cross-section (not to scale) of a single a-Si:H imaging pixel.

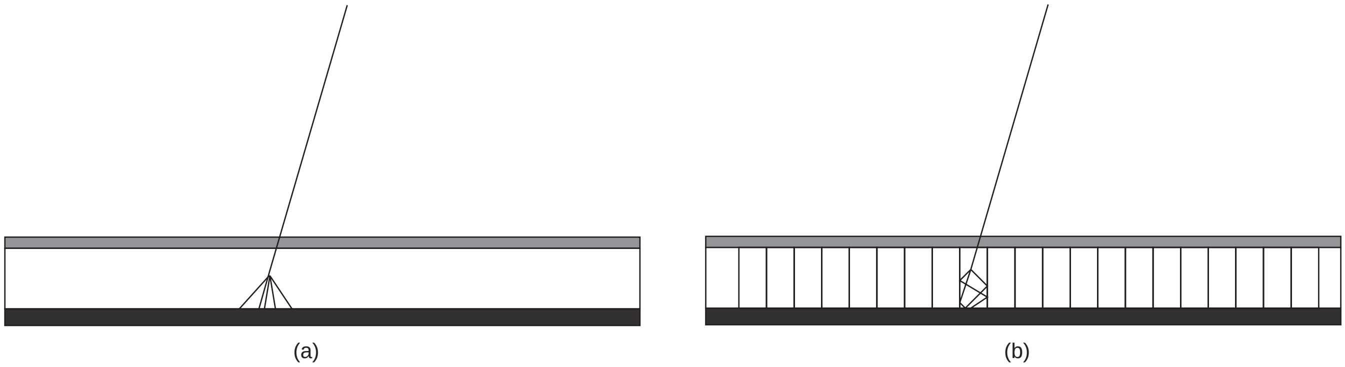


Figure 3. Illustration of EPID conversion screens used to trap optical light. Fig. 3a shows a regular Gd_2O_2S-screen with optical spread, while Fig. 3b shows a CsI screen which limits optical spread and increases detector resolution.

The resolution and efficiency of flat-panel imagers are theoretically superior to those from camera-based and LIC EPIDs. Research performed by Munro et al. (17) indicates that the amorphous silicon imager is X-ray quantum limited, and that the resolution is limited by the spread of the optical photons in the imager. The increase in quality and the compact size of TFT EPIDs means that they may be easily installed onto a linac using a robotic arm. Consequently, TFT EPIDs are now the only type of detectors commercially available. Efforts are underway to replace the current conversion screens, which usually are Gd_2O_2S phosphors, with CsI, which can be grown as single crystals the size of a pixel. The optical light is then trapped in a manner similar to an optical fiber, and therefore optical spread is eliminated, as shown in Fig. 3.

Figure 4 shows pelvic images taken with a camera-based detector, LIC, and flat-panel imager.

EPID APPLICATIONS

Replacement of Radiographic Film

Just as is common in radiology departments, electronic acquisition and management of imaging data is quickly becoming standard practice. Because of the fast pace of detector evolution in the past decade, EPID technology is facilitating hospital-wide imaging digitization. However, to understand why widespread implementation of EPID systems has not yet occurred, it is important to perform a brief cost analysis.

Compared to radiographic film-based portal imaging, up-front capital expenditures for EPID systems are about a factor of 5 larger (e.g., $25,000 vs. 125,000). However, on-going costs associated with an EPID system as compared to a radiographic film-based portal imaging program are much less. For example, regular purchasing of film and maintenance of a film processor (including silver harvesting) is not required with an EPID program. This cost-analysis makes EPID highly competitive given the digital direction facing modern health care.

With direct image digitization comes the ability to transmit data as required for telemedicine. Furthermore, imaging data storage and retrieval, as required for radiology picture archiving and communication systems (PACS), has additional advantages over conventional radiographic film storage. In radiation oncology departments, it is now commonplace for record-and-verify systems to be coupled to an electronic patient charting system. By storing the EPID images in this domain, many of the concerns for patient record keeping, retrieval, and preservation of treatment confidentiality are overcome.

Improvement of Patient Positioning

The original purpose of portal imaging is to reduce the incidence and extent of errors made during radiation treatment. The type of errors can be categorized as gross errors and stochastic errors. Examples of such errors are shown in Fig. 5. The gross errors occur only once and when detected can be removed from the treatment after one fraction. Stochastic errors contain a random component, which implies that the error changes from day to day. Errors and QA-problems can also introduce a systematic component hidden by the random component, which can only be corrected for if it's extent is known.

To reduce stochastic errors, there are two general methodologies, on- or off-line corrections. The most straightforward methodology uses on-line correction where an image is obtained in "localization mode" where minimal dose to the patient is applied. If a discrepancy is observed in the patient setup, and if this discrepancy is larger than a predetermined threshold or action level, efforts are taken to eliminate the error by changing the patient position or changing the treatment configuration. The aforementioned threshold is based on the precision to which position can be determined *and* with which the patient setup correction can be applied. Given the digital nature of EPID images, it is possible to increase the accuracy using computerized algorithms to objectively measure patient positioning with respect to the treatment field (18–21). This approach has not been widely adopted, mainly due to the perceived labor intensity and some medico-legal aspects. However, this may change with the advent of other on-line repositioning techniques (cf. ultrasound-based repositioning) and increased process automation.

Weekly port-filming is the standard procedure in the QA of external beam radiation therapy, which generally

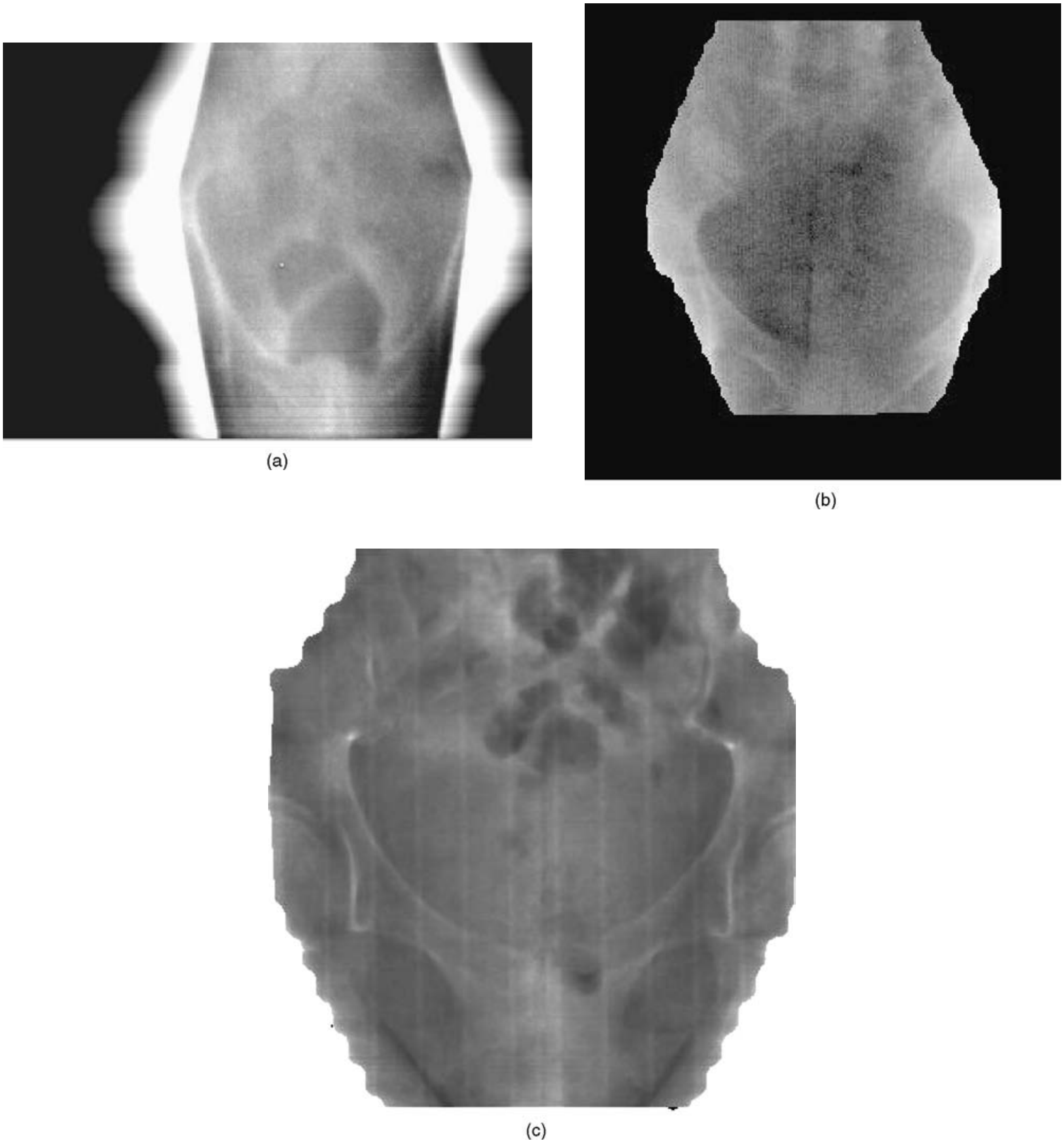


Figure 4. A comparison of pelvic images taken with three different types of EPIDs: 4a) a camera-based system, 4b) a liquid ionization chamber system, and 4c) a amorphous silicon (a-Si:H) flat panel imaging system.

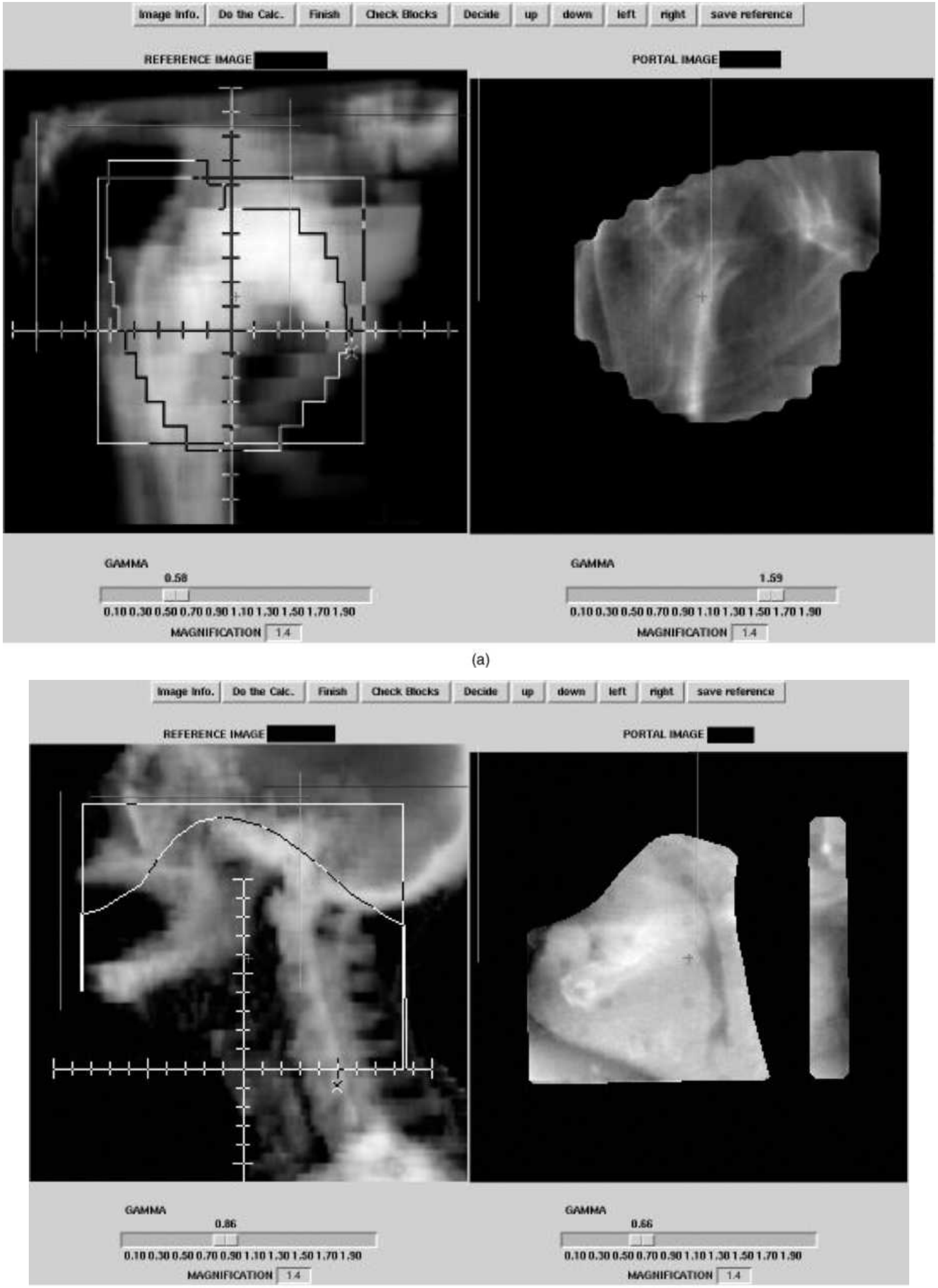
implies that a sample of 4 positions is taken out of a 20 fraction treatment. Errors are corrected after the first image. An interesting study by Valicente et al. (2), showed that this practice is suboptimal. The use of EPIDs allows us to obtain images in a more economical way. Most off-line correction strategies assume that the distribution formed by all consecutive errors is a normal distribution characterized by the mean error and the standard deviation calculated as in

Mean:

$$\langle x\rangle = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad (1)$$

Standard Deviation:

$$\sigma = \sqrt{\frac{1}{(N-1)}\sum_{i=1}^{N}(\langle x\rangle - x_i)^2}$$



GAMMA 0.58

MAGNIFICATION 1.4

GAMMA 1.59

MAGNIFICATION 1.4

(a)

GAMMA 0.86

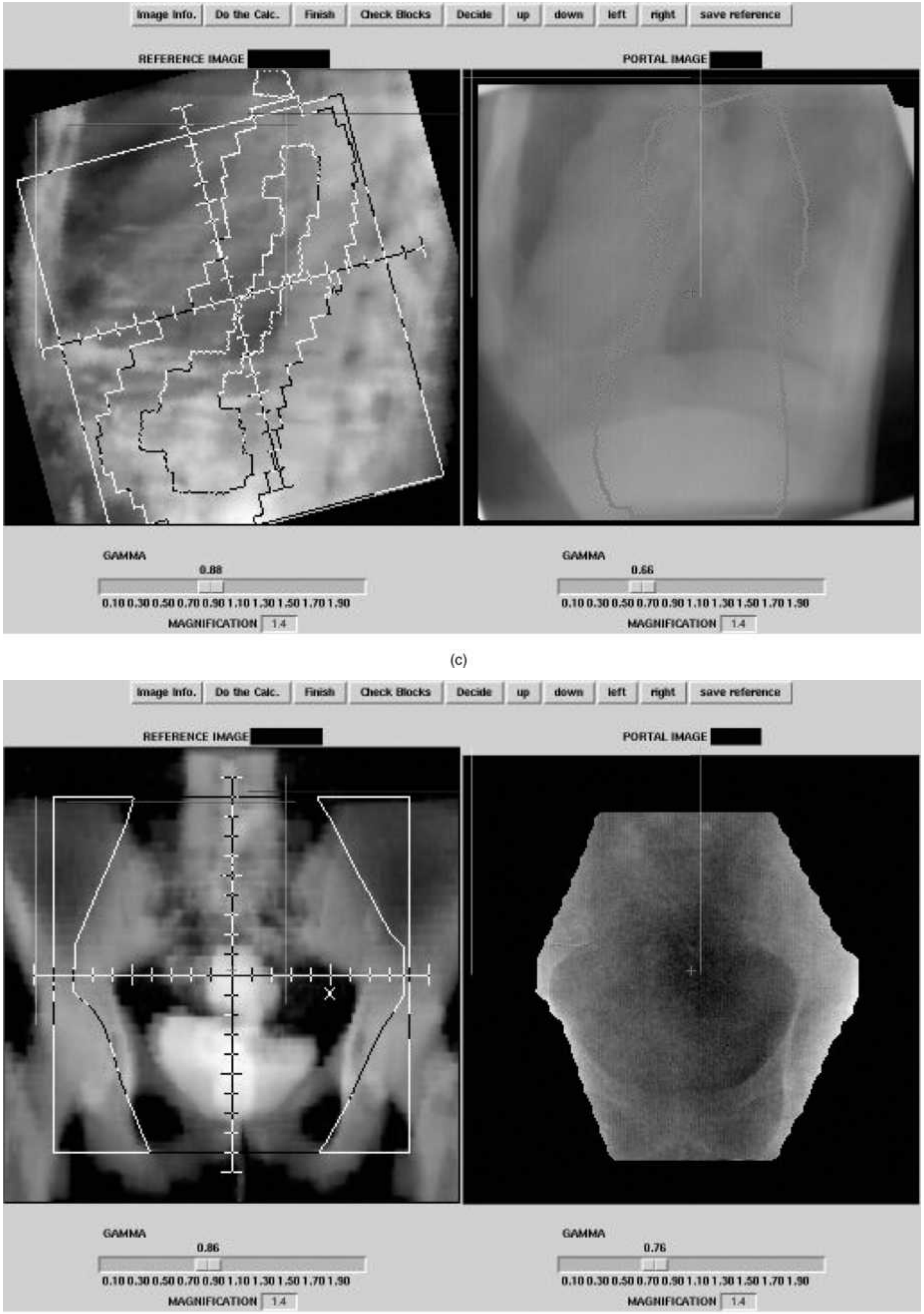
MAGNIFICATION 1.4

GAMMA 0.66

MAGNIFICATION 1.4

(b)

Figure 5. Examples of setup errors. Reference images on the right are digitally reconstructed radiographs with the correct setup. On the left are the measured portal images. (a) Faulty collimator angle. (b) Wrong blocking used. (c) Wrong MLC file. (d) Patient positioning error.



GAMMA 0.88

MAGNIFICATION 1.4

GAMMA 0.66

MAGNIFICATION 1.4

(c)

GAMMA 0.86

MAGNIFICATION 1.4

GAMMA 0.76

MAGNIFICATION 1.4

(d)

Figure 5. (*Continued*)

By repeated sampling of the distribution (e.g., taking port films), the strategy estimates the value of $\langle x \rangle$ as close as possible and corrects for the error, which will reduce the systematic error in the treatment. Several groups have studied the implications of this strategy and its variations. The most successful approach seems to be the following strategy: The set-up variations are recorded and averaged. This is compared to an action level that depends on how many samples have been taken already (the level shrinks as the the amount of information on the systematic error increases). These studies showed that systematic errors could be reduced to ~2 mm (22).

Organ Motion

One of the major reasons for use of EPIDs is the fact that the patients anatomy and position vary from those used for treatment planning purposes. The factors involving this variation are

- Patient movement.
- Patient positioning inaccuracies.
- Organ motion.

Any of these factors will influence the actual dose distribution to be different from that obtained using treatment planning. It is straightforward to correct for the first two problems using portal imaging as the patients position is typically well-characterized using bony anatomy. An excellent compilation on the incidence, extent, and repercussions of organ movement was performed by Langen and Jones (23). Efforts to incorporate organ movement during radiotherapy treatment planning involves enlarging the target to be treated. Sophisticated algorithms that calculate the extent of these enlargements were developed independently by Stroom et al. (24) and by van Herk et al. (25). The general framework for this enlargement is given in ICRU 50 and ICRU 60 (26,27). In these reports the gross target volume (GTV) is defined as the volume containing demonstrated tumor. A margin is added to the GTV to account for suspected microscopic tumor involvement, and is defined as the clinical target volume (CTV). Finally, the planning target volume (PTV) is defined by the CTV and an additional margin to allow for geometrical variations such as patient movement, positioning errors, and organ motion. The margins added to GTV and CTV can substantially increase the PTV since the margins are applied in three dimensions. Because of volume effects of radiation therapy, there is a tendency to minimize the PTV by increasing the precision of the treatment. As explained above, EPIDs are able to minimize uncertainties caused by patient motion and positioning errors using off- or on-line correction strategies.

Except in a few cases like in lung or where air pockets are present (24,28), the target is virtually indiscernible using X rays. To solve this problem, other modalities, like CT (29) and ultrasound (30,31), have been used to determine the position of the organ. The EPIDs can also be used to image the position of organs if radioopaque markers are implanted. The markers need to be of high density and migration needs to be limited. The efficacy and feasibility of using markers with EPIDs was studied by Balter et al. (32), application of the use of markers have been extensively studied by Pouliot and co-workers (33). The use of markers is becoming more popular and EPID systems are being augmented with software to detect markers as well as perform the requisite positional calculations.

Quality Assurance and *In Vivo* Dosimetry

In 2001, Task Group No. 58 of the American Association of Physicists in Medicine Radiation Therapy Committee issued a protocol to define the standard-of-care for performing EPID QA on a daily, monthly, and annual basis (34). In this protocol, a quality assurance program is proposed where daily checks of EPID system performance, image quality, and safety interlocks are performed by a radiation therapy technologist. In addition to reviewing results and independently performing checks conducted by the technologist, a medical physicist should conduct the following checks on a monthly basis: perform constancy check of SNR, resolution, and localization; inspect images for artifacts; do a mechanical inspection of all EPID components; and maintain the computer system. The annual QA tasks are also performed by the medical physicist, and include all of the above tasks plus a full check of the EPID geometric localization accuracy. By performing these QA tasks, the radiotherapy department may be reasonably assured of a reliable EPID system for clinical use.

The QA tests typically utilize a vendor-supplied phantom designed to facilitate evaluation of the aforementioned tasks. Since clinical linacs are typically dual energy (e.g., 6 and 15 MV) in design, tests are applied to both photon energies. As expected, the lower photon energy will demonstrate improved image quality (e.g., SNR and spatial resolution). Since many EPID systems utilize sophisticated computer software utilities, testing of this software in a realistic setting is an integral component of the EPID quality management program. As can be expected with tests that are subjective in nature, it is recommended that multiple users be employed to evaluate the subjective criteria so as to minimize user bias.

In addition to the aforementioned advantages, an EPID system permits unique opportunities of radiotherapy treatment QA. Treatment fields are typically blocked with beam modifiers to account for irregularities in patient shape. These beam modifiers include compensating materials when minor changes are required, or beam wedges when gross changes are required. Because modern linacs have features like dynamic wedges and dynamic multileaf collimators, the nonintegral approach that EPIDs offer over radiographic film (i.e., the ability to obtain several images at different stages during a dynamic process) permits continued high-quality QA. Due to the electronic nature of EPID measurement of the photon energy fluence exiting a patient, one can perform exit dosimetry and quantitative comparisons with treatment planning intentions (35). However, these efforts are currently research driven, and widespread clinical implementation may not be expected for a few years.

BIBLIOGRAPHY

1. AAPM report No. 24: Radiotherapy portal imaging quality, 1987.
2. Valicenti RK, Michalski JM, Bosch WR, Gerber R, Graham MV, Cheng A, Purdy JA, Perez CA. Is weekly port filming adequate for verifying patient postion in modern radiation therapy? Int J Rad Oncol Biol Phys 1994;30(2):431–438.
3. Moseley J, Munro P. Display equalization: A new display method for portal images. Med Phys 1993;20(1):99–102.
4. Van den Heuvel F, Han I, Chungbin S, Strowbridge A, Tekyi-Mensa S, Ragan D. Development and clinical implementation of an enhanced display algorithm for use in networked electronic portal imaging. Int J Rad Oncol Biol Phys 1999;45:1041–1053.
5. Jaffray DA, Batista JJ, Fenster A, Munro P. X-ray scatter in megavoltage transmission radiography: Physical characteristics and influence on image quality. Med Phys 1994;21(1):45–60.
6. Bailey NA, Horn RA, Kamp TD. Fluoroscopic visualization of megavoltage therapeutic x-ray beams. Int J Radiat Oncol Biol Phys 1980;6:935–939.
7. Leszczynski KW, Shalev S, Cosby S. A digital video system for on-line portal verification. Medical Imaging IV: Image Formation. SPIE 1990;1231:401–405.
8. Leong J. Use of digital fluorscopy as an on-line verification device in radiation therapy. Phys Med Biol 1986;31:985–992.
9. Munro P, Rawlinson JA, Fenster A. A digital fluoroscopic imaging device for radiotherapy localization. Int J Rad Oncol Biol Phys 1990;18:641–649.
10. Visser AG, Huizenga H, Althof VGM, Swanenburg BN. Performance of a prototype fluoroscopic radiotherapy imaging system. Int J Rad Oncol Biol Phys 1990;18:43–50.
11. Wong JW, Slessinger ED, Hermes RE, Offutt CJ, Roy T, Vannier MW. Portal dose images I: Quantitative treatment plan verification. Int J Rad Oncol Biol Phys 1990;18:1455–1463.
12. Wickman GA. A liquid filled ionisation chamber with high spatial reslution. Phys Med Biol 1974;19:66–72.
13. Meertens H, van Herk M, Weeda J. A liquid ionisation detector for digitital radiography of therapeutic megavoltage photon beams. Phys Med Biol 1985;30:313–321.
14. Van Herk M, Meertens H. A matrix ionization chamber imaging device for on-line patient set-up verification during radiotherapy. Radiother Oncol 11:369–378.
15. Morton EJ, Swindell W, Evans PM. A linear array, scintillation crystal-photodiode detector for megavoltage imaging. Med Phys 1991;18:681–691.
16. Lam KS, Partowmah M, Lam WC. An on-line electronic portal imaging system for external beam radiotherapy. Br J Radiol 1986;59:1007–1013.
17. Munro P, Bouius DC. X-ray quantum limited portal imaging using amorphous silicon flat-panel arrays. Med Phys 1998;25(5):689–702.
18. De Neve W, Van den Heuvel F, De Beukeleer M, Coghe M, Verellen D, Thon L, De Roover P, Roelstraete A, Storme G. Interactive use of on-line portal imaging in pelvic radiation. Int J Rad Oncol Biol Phys 1993;25:517–524.
19. Van den Heuvel F, De Neve W, Verellen D, Coghe M, Coen V, Storme G. Clinical implementation of an objective computer-aided protocol for intervention in intra-treatment correction using electronic portal imaging. Radiother Oncol 1995;35:232–239.
20. Balter JM, Pelizarri CA, Chen GTY. Correlation of projection radiographs in radiation therapy using open curve segments and points. Med Phys Mar.–Apr. 1992;19(2): 329–334.
21. Bijhold J, Lebesque JV, Hart AAM, Vijlbrief RE. Maximizing setup accuracy using portal images as applied to a conformal boost technique for prostate cancer. Radiother Oncol 1992;24:261–271.
22. Bel A, Vos PH, Rodrigus PTR, Creutzberg CL, Visser AG, Stroom JC, Lebesque JV. High-precision prostate cancer irradiation by clinical application of an offine patient setup verification procedure, using portal imaging. Int J Rad Oncol Biol Phys 1996;35(2).
23. Langen KM, Jones DT. Organ motion and its management. Int J Radiat Oncol Biol Phys May 1 2001;50(1):265–278.
24. Stroom JC, Boer de HC, Huizenga H, Visser AG. Inclusion of geometrical uncertainties in radiotherapy treatment planning by means of coverage probability. Int J Radiat Oncol Biol Phys Mar 1 1999;43(4):905–919.
25. van Herk M, Remeijer P, Lebesque JV. Inclusion of geometric uncertainties in treatment plan evaluation. Int J Radiat Oncol Biol Phys Apr 1 2002;52(5):1407–1422.
26. ICRU and International Commission on Radiation Units and Measurements. Prescribing, recording and reporting photon beam therapy. ICRU Report 1993; 50.
27. ICRU and International Commission on Radiation Units and Measurements. Prescribing, recording and reporting photon beam therapy, Supplement to ICRU 50. ICRU Report 1999; 62.
28. Erridge SC, Seppenwoolde Y, Muller SH, van Herk M, De Jaeger K, Belderbos JSA, Boersma LJ, Lebesque JV. Portal imaging to assess set-up errors, tumor motion and tumor shrinkage during conformal radiotherapy of non-small cell lung cancer. Radiother Oncol 2003;66(1):75–85.
29. Jaffray DA, Drake DG, Moreau M, Martinez AA, Wong JW. Radiographic and tomographic imaging system integrated into a medical linear accelerator for localization of bone and soft-tissue targets. Int J Radiat Oncol Biol Phys Oct 1 1999;45(3):773–789.
30. Lattanzi J, McNeeley S, Hanlon A, Schultheiss TE, Hanks GE. Ultrasound-based stereotactic guidance of precision conformal external beam radiation therapy in clinically localized prostate cancer. Urology Jan 2000;55(1): 73–78.
31. Serago CF, Chungbin SJ, Buskirk SJ, Ezzell GA, Collie AC, Vora SA. Initial experience with ultrasound localization for positioning prostate cancer patients for external beam radiotherapy. Int J Radiat Oncol Biol Phys Aug 1 2002;53(5):1130–1138.
32. Balter JM, Sandler HM, Lam K, Bree RL, Lichter AS, ten Haken RK. Measurement of prostate movement over the course of routine radiotherapy using implanted markers. Int J Rad Oncol Biol Phys 1995;31(1):113–118.
33. Vigneault E, Pouliot J, Laverdiere J, Roy J, Dorion M. Electronic portal imaging device detection of radioopaque markers for the evaluation of prostate position during megavoltage irradiation: A clinical study. Int J Radiat Oncol Biol Phys Jan 1 1997;37(1): 205–212.
34. Herman MG, Balter JM, Jaffray DA, McGee KP, Munro P, Shalev S, van Herk M, Wong JW. Clinical use of electronic portal imaging: report of radiation therapy committee task group 58. Med Phys May 2001;28(5):712–737.
35. Boellaard R, Van Herk M, Mijnheer BJ. A convolution model to convert transmission dose images to exit dose distributions. Med Phys 1997;24(2):189–199.

See also COMPUTED TOMOGRAPHY; MAGNETIC RESONANCE IMAGING; PHOTOGRAPHY, MEDICAL; POSITRON EMISSION TOMOGRAPHY; ULTRASONIC IMAGING.

IMMUNOLOGICALLY SENSITIVE FIELD–EFFECT TRANSISTORS

Emmanuel S. Zachariah
University of New Jersey
New Brunswick, New Jersey

P. Gopalakrishnakone
National University of Singapore
Singapore

Pavel Neuzil
Institute of Bioengineering
and Nanotechnology
Singapore

INTRODUCTION

Diagnostics as a whole represent a large, well-established, and continually expanding market. Methods for the selective determination of analytes in biological fluids, such as blood and urine, are important. When a foreign substance (antigen) invades the human body, the immune system produces antibodies that interact with the antigen. Such a recognizing process involves the formation of an immunocomplex based on interactions between the immunospecies. The recognition is specific for the antibody-antigen system and, thus, for the measurement of antigen concentration (1,2). This determination is of importance for diagnosis because the antigens can be viruses, bacteria that are involved in many human illnesses, such as cancer and AIDS. The analytes detected and measured have also included many other medical diagnostic molecules such as hormones, clinical disease biomarkers, drugs, and environment pollutants such as pesticides. Antibody diversity is so great that virtually any biomolecule can be recognized (3). The range of analyte concentrations encountered is extremely large, from greater than $10^{-3}\,M$ for species such as glucose and cholesterol and to less than $10^{-12}\,M$ for certain drugs and hormones (4). It is for the detection of these low level analytes that the application of immunological techniques is essential.

An immunoassay is a multistep diagnostic test based on the recognition and binding of the analyte by the antibody. Most immunoassay techniques are based on the separation of free and bound immunospecies (5). In these techniques, one of the immunoagents (antibody or antigen) is immobilized on a solid phase. The solid phase facilitates the separation and washing steps required to differentiate bound and free fractions of the label. Quantification of a bound immunoagent is conducted by using labels covalently bound to the immunoagent with specific properties suitable for detection. The most common labels are radioactive markers, enzymes, and fluorescent labels. For many of the nonisotopic labels, the reagents have been designed such that binding of labeled antigen to antibody in some way modulates the activity of the label, resulting in a homogenous immunoassay without the need for a separation step. The most familiar type of enzyme immunoassay in clinical analysis is known as enzyme-linked immunosorbent assay (ELISA) (6). Different schemes of enzyme immunoassay exist, and, in clinical laboratory practice, the most popular are the “Sandwich” method for large analytes, and competitive binding immunoassay methods for the determination “haptens” (low molecular weight analytes).

The advent of biosensor technology, with the possibility of direct monitoring of immunoreactions, provides opportunity to gain new insight into antigen-antibody reaction kinetics and create rapid assay devices with wider applications. A biosensor is composed of (*1*) a biochemical receptor, which uses biosubstances such as enzymes, antibodies, or microbes to detect an analyte, (*2*) a transducer, which transforms changes in physical or chemical value accompanying the reaction into a measurable response, most often in the form of electrical signal (7–9). The term immunosensor is used when antibodies are immobilized to recognize their appropriate antigens (or vice versa) (10). Immunosensors possess several unique features, such as compact size, simplicity of use, one-step reagentless analysis, and absence of radioactivity, which make them attractive alternatives to conventional immunoassay techniques. Immunosensors can be divided, in principle, into two categories: nonlabeled and labeled (11). Nonlabeled or direct-acting immunosensors are designed in a way that the immunocomplex (i.e., the antibody-antigen complex) is directly determined by measuring physical changes induced by the formation of the complex. In contrast, labeled or indirect-sensing immunosensors have incorporated a sensitively detectable label. The immunocomplex is thus entirely determined through measurement of the label. In order to determine an antigen, the corresponding antibody is immobilized on the membrane matrix, which is held on an amperometric-or potentiometric-sensing transducer used to measure the rate of the enzymatic reaction (12–14).

Of the electrochemical technologies for biosensors, the Ion-Sensitive Field Effect Transistor (ISFET) has been the center of special attention as a transducer. ISFETs were introduced by Bergveld in 1970 (15), and were the first type of this class of sensor in which a chemically sensitive layer was integrated with solid-state electronics. A field effect transistor (FET) can be considered as a charge-sensitive device (i.e., any change in the excess interfacial charge at the outer insulator surface will be mirrored by an equal and opposite charge change in the inversion layer of the FET). By excluding the gate metal in a FET and using a pH-sensitive gate insulator, a pH-sensitive FET was constructed (16,17). After the invention of the ISFET, many different types of FET-based sensors have been presented. The application of enzymes as the selecting agent in ISFET-based sensing systems leads to the development of highly sensitive sensors. Such enzyme-modified ISFETs (EnFETs) can, in principle, be constructed with any enzyme that produces a change in pH on conversion of the concerning substrate (18). By combining the ISFET with a membrane that contains a biological substance, like an antibody, the sensor can detect a specific antigen (19). The ISFET immunosensors or Immunologically sensitive FETs (IMFETs) have several advantages over the conventional enzyme immunoassay. The ISFET could be mass-produced by an integrated circuits (IC) process, which makes it very small and economical. An electric circuit can be integrated on the same chip. The biosensor platform finds many applications in various fields, such as medical diagnostics, fermentation process control, and environmental monitoring.

THEORY

In order to understand the operation of the IMFET, one must trace its origins back to the ISFET or ChemFET (Fig.1). The latter devices have been described in depth elsewhere (20–22). A packaged ISFET is shown in Fig. 2 (23). ISFETs and ChemFETs have, in turn, evolved from the Metal Oxide Semiconductor Field Effect Transistor (MOSFET), currently the most popular active device in the entire semiconductor industry. It is a unipolar device, where the current is given by the flow of majority carriers, either holes in PMOS type or electrons in NMOS type. The operation of the MOSFET can be considered as a resistor controlled by the status of a gate region, so-called MIS structure. It is a sandwich consisting of a stacked-gate metal layer, an insulator (typically silicon oxide), and a semiconductor. Assume a low-level doped p-type (NMOS device). Three different states of charge distribution can occur, depending on the voltage V_g, applied between the metal and a semiconductor. A negative value of V_g causes positive holes to accumulate at the semiconductor-insulator interface. A positive value of V_g of a low magnitude leads to the “depletion” condition in which mobile holes are driven away from the interface, resulting in a negative charge of low density due to the presence of immobile acceptor atoms. Finally, if the V_g exceeds a certain threshold voltage (V_{th}), electrons accumulate at the semiconductor-insulator interface at a density greater that the hole density, a situation opposite to that normally found with p-type semiconductors. This depletion of mobile charge carriers followed by surface inversion is known as the “field effect.” It forms an electrically conductive channel between two other terminals, a source and a drain (see Figure 1a). The drain current I_d through the transistor is a functions of drain and gate voltage. Without surface inversion (i.e., $V_g < V_{th}$,) the drain current is negligible, because the drain-to-substrate PN junction is reverse biased.

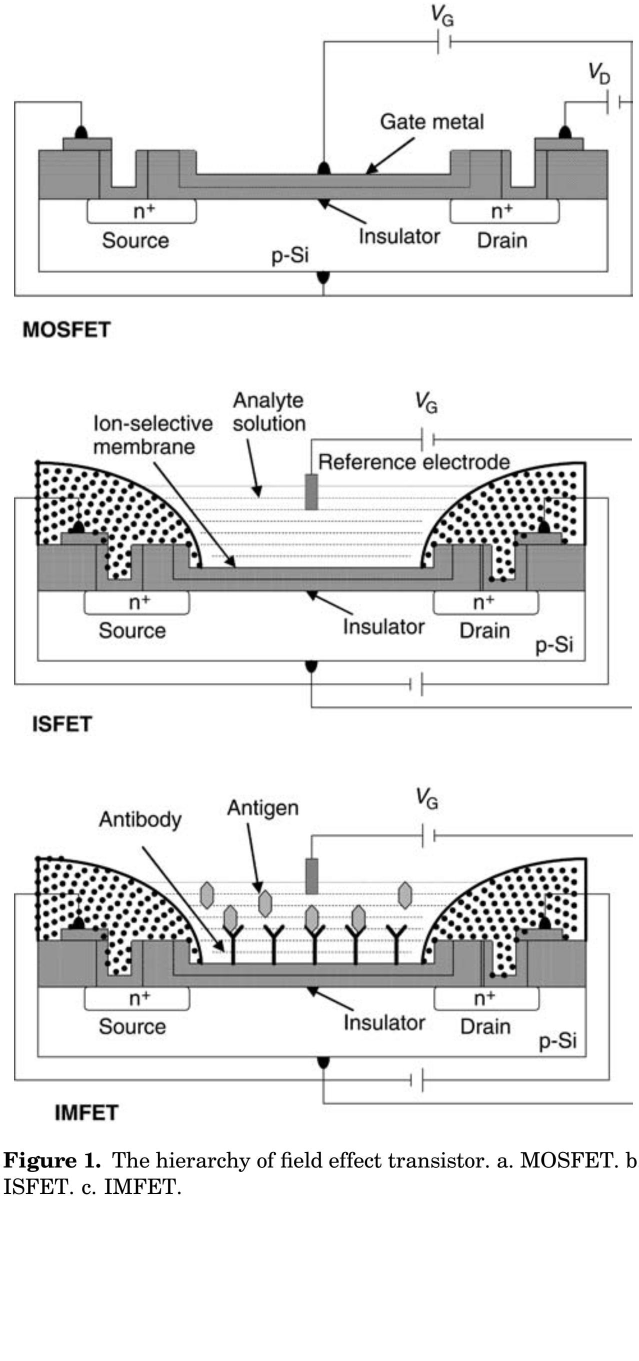


Figure 1. The hierarchy of field effect transistor. a. MOSFET. b. ISFET. c. IMFET.

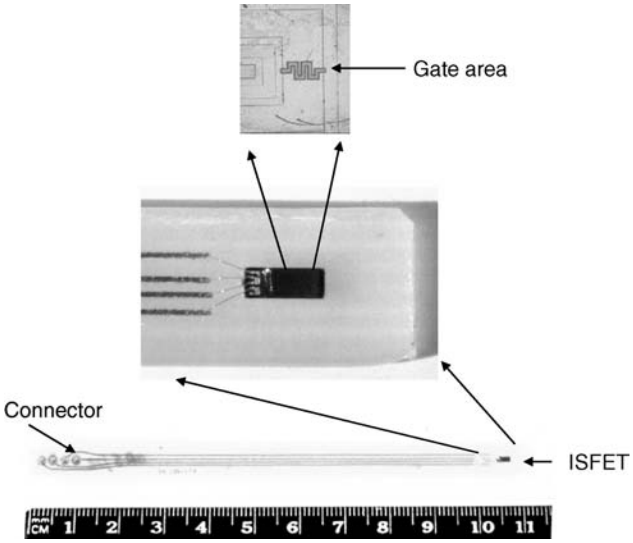


Figure 2. Photomicrograph of an ISFET device packaged on PCB (23).

The MOSFET and its descendants are charge-controlled devices. In analytical applications (e.g., ISFETs, ChemFETs, and IMFETs), the change in charge density is brought about by adsorption of one or more species present in the solution onto the FET structure. In the ISFET, the gate metal is replaced with a conventional reference electrode (Ag/AgCl or Hg/Hg_2Cl_2), a solution containing an ionic species of interest, and an electroactive material (membrane) capable of selective ion exchange with the analyte (Fig. 1b), which is an example of a nonpolarizable interface, that is, reversible charge transfer occurs between the solution and the membrane. The analyte generates a Nernst potential at the membrane-solution interface, which then modulates the drain current analogous to the manner in which changing the externally applied voltage does for the MOSFET.

Direct-Acting (Label-Free) IMFET

The structure of the direct-acting IMFET is similar to that of the ISFET, except that the solution-membrane interface is polarized rather than unpolarized. If the solution-membrane interface of the ISFET is ideally polarized (i.e., charge cannot cross the interface), then the ISFET can measure the adsorption of charged species at the interface as shown below. As antibodies, antigens, and proteins are generally electrically charged molecules, the polarized ISFET could be used to monitor their nonspecific adsorption at the solution-membrane interface. To render the polarized ISFET selective for a given antigen and thus create the so-called IMFET, the specific antibody for that antigen has to be immobilized on the surface of the ISFET (see Fig. 1c). The adsorption of this antigen would then be specifically enhanced over other molecules in the solution and the signal measured by the ISFET would be mostly due to the adsorption of that particular antigen. The ISFET interacts with the analyte through an ion-exchange mechanism, whereas the IMFET interaction is based on the antigen-antibody reaction.

This design for the measurement of the adsorption of charged molecules is practicable only if charge cannot cross the interface, which, thus, acts as an ideal capacitor. As will be seen, failure to achieve a perfectly polarizable interface has a detrimental effect on the specificity of the IMFET. Few reports exist on direct-acting IMFETs; a brief analysis on the work of Janata research group will be presented here (24–26). The capacitance of a polarized interface is described by electrical double-layer theory and is usually modeled as a series combination of two capacitors, C_G and C_H, where C_G is the capacitance of the diffuse Gouy–Chapman part of the double layer and C_H is the capacitance of the Helmholtz part of the double layer (27). The total capacitance, C_{dl}, is therefore

$$1/C_{dl} = 1/C_G + 1/C_H \tag{1}$$

The electrical circuit through the gate of an ISFET with an ideally polarized interface can be modeled, therefore, as a series combination of C_G, C_H, and C_0, as drawn in Fig. 3, where C_0 is the capacitance of the insulator. A gate voltage V_G is applied through a reference electrode between the solution and the semiconductor. The process of adsorption of charged molecules can be modeled as the transfer of a quantity of charge from the solution to the surface of the transistor as would occur if the switch were closed for a short time period allowing the current source to transfer the charge. As adsorption occurs, the charge on each plate of the capacitors will change to accommodate the new charge balance. The charge change on capacitor C_0 is the quantity of interest as it represents the charge in the inversion layer of the FET, Q_i, and will affect the drain current of the transistor, which can be directly measured. If a quantity of charge, Q_{ads}, is transferred by the adsorption of charged molecules, then the charge change on C_0, Q_i, can be represented by

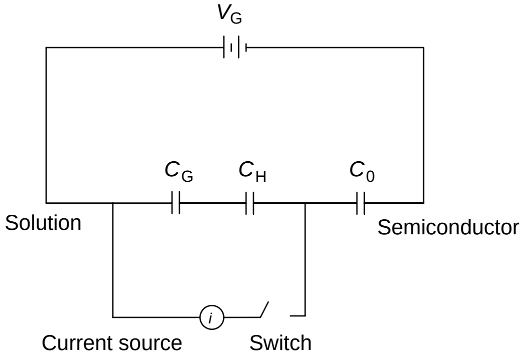


Figure 3. Electrical model for the measurement of charge adsorption with the CHEMFET.

$$Q_i = Q_{ads}\{C_0/(C_0 + C_{dl})\} \tag{2}$$

Hence, only a fraction of the adsorbed charge will be mirrored in the transistor. When adsorption occurs, because electroneutrality must be observed in the system, an equal quantity of the opposite charge must either enter the inversion layer of the FET or enter the double layer from the solution. Equation 2 predicts that part of the image charge will come from the solution as ions entering the double layer with the adsorbing molecules. This fraction of charge, which is mirrored in the inversion layer of the FET, will be defined as β, and it is defined as

$$\beta = Q_i/Q_{ads} = C_0/(C_0 + C_{dl}) \tag{3}$$

According to this model, only 0.3% of the charge on the adsorbing molecules will be mirrored in the inversion layer of the FET. The authors conservatively estimated β to be 10^{-4}. Considering the I_d current as a function of the potential at the solution-membrane interface, it is clear that a relationship between the adsorbed charge and interfacial potential, $\Phi_{Sol\text{-}mem}$, is necessary to describe the chemical response of the IMFET. This potential is merely the charge change induced in the inversion layer divided by the insulator capacitance:

$$\Phi_{Sol-mem} = Q_i/C_0 = \beta Q_{ads}/C_0 \tag{4}$$

Substitution of this expression in to Equations 5 and 6 yields the response equations for the polarized ISFET. The authors derived the following expressions for the polarized ChemFET relating to Q_i to the observed parameter, the drain current (I_d):

$$I_d = \frac{\mu n W C_0}{L}\left(V_g - V_t - E_r - \phi_{sol-mem} - \frac{V_d}{2}\right)V_d \qquad V_d < V_{dsat} \tag{5}$$

and

$$I_d = \frac{\mu n W C_0}{2L}(V_g - V_t - E_r - \phi_{sol-mem})^2 \quad V_d > V_{dsat} \tag{6}$$

where W is the width of the source-drain conducting channel, μ_n is the effective electron mobility in the channel, C_0 isthe capacitance per unit area of the gate insulator, L is the channel length, V_d is the drain-to-source voltage, V_g is the applied gate voltage, V_t is the threshold voltage (for surface inversion), and E_r is the potential of the reference electrode.

The antibody-antigen binding reaction in its simplest form can be expressed in terms of the following biomolecular reaction:

$$\mathrm{Ab} + \mathrm{Ag} \rightleftarrows \mathrm{AbAg}$$

where Ab is the antibody, Ag is the antigen, and AbAg is the complex. The reaction is characterized by the equilibrium constant K,

$$K = [\mathrm{AbAg}]/[\mathrm{Ab}][\mathrm{Ag}] \tag{7}$$

The total charge change at the interface due to the binding, Q_{i}, can be shown to be

$$Q_{\mathrm{i}} = \beta Q_{\mathrm{ads}} = \beta z F \frac{K[\mathrm{Ag}][\mathrm{S}]}{1 + [\mathrm{Ag}]} \tag{8}$$

where z is the ionic charge of the antigen and [S] is the surface concentration of binding sites (the surface concentration of immobilized antibodies before binding). Substitution of this expression into Equation 4 yields

$$\Phi_{\mathrm{Sol-mem}} = \frac{\beta z F K[\mathrm{Ag}][\mathrm{S}]}{C_0(1 + [\mathrm{Ag}])} \tag{9}$$

From Equation 9, the limit and range of the detection for the IMFET can be predicted. Assume that the equilibrium constant is in typical range from 10^5 to 10^9 (28), which gives a value of $\beta = 10^{-4}$. If the antibodies are immobilized with a surface concentration of 1 molecule per 10 nm^2 and the charge on an antigen is five electronic charges of an antibody, the IMFET's detection limit would be in the range of $10^{-7} - 10^{-11}$M of concentration antibody concentration. The antigen concentration that gives 90% surface coverage can similarly be calculated to be in the range of $10^{-4} - 10^{-8}$ M. Similar equations can be derived for the case where the antigen is immobilized at the interface rather than the antibody. However, it has been argued by many researchers that a static measurement concerning the presence of a protein layer on an electrode is difficult, because the charged groups are, of course, neutralized by surrounding counter ions (29). In order to avoid interference from other charged species present in the solution, the substrate for immobilization should preferably be inert and nonionic (24–30), which in aqueous solutions implies a hydrophobic surface (31). Ideal conditions that are required in this coherence are a truly capacitive interface at which the immunological binding sites can be immobilized, a nearly complete antibody coverage, highly charged antigens, and a low ionic strength.

Schasfoort et al. (32) extensively studied the requirements for the construction of IMFET, which would operate on the direct potentiometric sensing of protein charges. The charge redistribution around immobilized proteins at the insulator-solution interface can be described by the double-layer theory (33). On adsorption, the diffuse layer of counter ions around the protein charges may overlap with the diffuse layer of the electrolyte-insulator interface. The thickness of diffuse-charge layers is described by the Debye theory (34) and defined by the distance where the electrostatic field has dropped to 1/ e of its initial value:

$$\kappa^{-1} = \left(\frac{\varepsilon_0 \varepsilon k T}{2q^2 I}\right)^{1/2}$$

where κ^{-1} is the Debye length, q the elementary change, k Boltzmann's constant, T absolute temperature, ε_0 the permittivity of vacuum, ε the dielectric constant, and $I = 1/2\sum c_i z_i^2$ represents the ionic strength in which c_i is the concentration of ion i with valency z (for 1-1 salt, I can be replaced by c).

It can be seen from the equation that the Debye length is strongly dependent on the ionic strength of the solution; more precisely, the Debye length is inversely proportional to the square root of the ionic strength. Therefore, one can expect that the chance of overlapping of the double layers of the substrate-solution interface and the adsorbed proteins can be substantial only if low electrolyte concentrations are used, owing to the dimensions of the proteins (Fig. 4). In a physiological salt solution, the Debye length is limited to ca. 0.8 nm. It is obvious that only charge density changes that occur within the order of a Debye length of the ISFET surface can be detected. With the macromolecules, such as protein, the dimensions are much larger (about 10 nm) than those of the double layer of the electrolyte-insulator interface, which means that, in such a case, most of the protein charge will be at a distance greater than the Debye length from the surface. If, moreover, on top of a monolayer of antibody molecules a second layer on antigens in coupled, it is obvious that the chance of overlap of the diffuse layers of antigens with electrolyte-substrate interface will decrease even more. At high ionic strength, the additional charges of the antigen are nearly always located far outside the diffuse layer at the ISFET surface and pure electrostatic detection of these antigenic charges, therefore, is impossible. In addition, a theoretical approach is

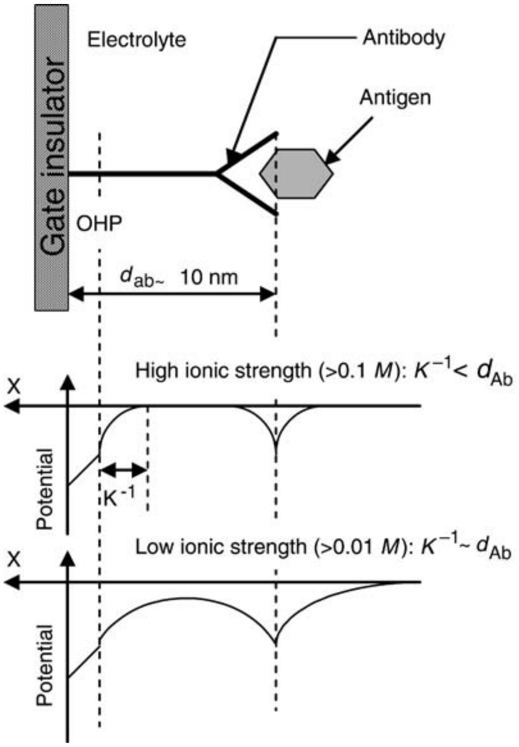

Gate insulator
Electrolyte
Antibody
Antigen
OHP
$d_{\mathrm{ab}\sim}$ 10 nm
X
High ionic strength (>0.1 M): $K^{-1} < d_{\mathrm{Ab}}$
Potential
K^{-1}
X
Low ionic strength (>0.01 M): $K^{-1} \sim d_{\mathrm{Ab}}$
Potential

Figure 4. Schematic representation of the potential distribution in a direct-acting IMFET. K^{-1} is the Debye length; d_{Ab}, dimension of macromolecule (e.g., antibody).

given based on the Donnan equilibrium description, which provides an insight into the potential and ion distribution in the protein layer on the IMFET (32). It is shown that the Donnan potential and the internal pH shift, induced by the protein charges, compensate each other to a greater extent. If the ISFET shows Nernstian behavior, it can be concluded that a direct detection of protein charge is impossible. In order to construct an IMFET, a reference FET or ISFET with a low sensitivity would satisfy the detection of the partially compensated Donnan potential in the presence of an adsorbed protein layer. However, the application of such as IMFET is limited to samples with low ionic strength.

An alternative, indirect approach is proposed by Schasfoort et al. (35,36) for the detection of an immunological reaction taking place in a membrane, which covers the gate area of an ISFET (Figs. 5a and 5b). The protein layer on the gate is exposed to pulse-wise increases in electrolyte concentration. As a result, ions will diffuse into the protein layer and, because of a different mobility of anions and cations, transients in potential will occur at the protein-membrane solution interface. The ISFET, being a voltage-sensitive device, is suitable for the measurement of these transients. As the mobility of ions is a function of the charge density in the protein membrane, changes in the charge density will influence the size and direction of the transients. By exposing the ISFET to a pH gradient and a continuous series of ion concentration pulses, the isolectric point of the protein layer can be detected and, thus, changes as the result of an immunological reaction. When a membrane separates two compartments with different electrolyte concentrations, a potential gradient can be measured. The different diffusion rates of anions and cations through the membrane set up a static membrane potential, which can be expressed by the Nernst–Planck equation (33):

$$\phi_m = \frac{RT}{F}.U.In\frac{a_2}{a_1} \quad U = \frac{D_+ - D_-}{D_+ + D_-}$$

where ϕ_m = the membrane potential, RT and F have their common meaning, U = the ratio of the diffusion coefficients (D_+ and D_-) of cations and anions, and a_1 and a_2 are the electrolyte activities in the respective compartments. The ion-step method is further developed by Schasfoort and Eijsma (37), and a detailed theoretical understanding of the ion-step response has been presented by Eijiki et al. (38). Recently, an impedance spectroscopy method was used tocharacterize immobilized protein layers on the gate of an ISFET and to detect an antigen-antibody recognition event (39).

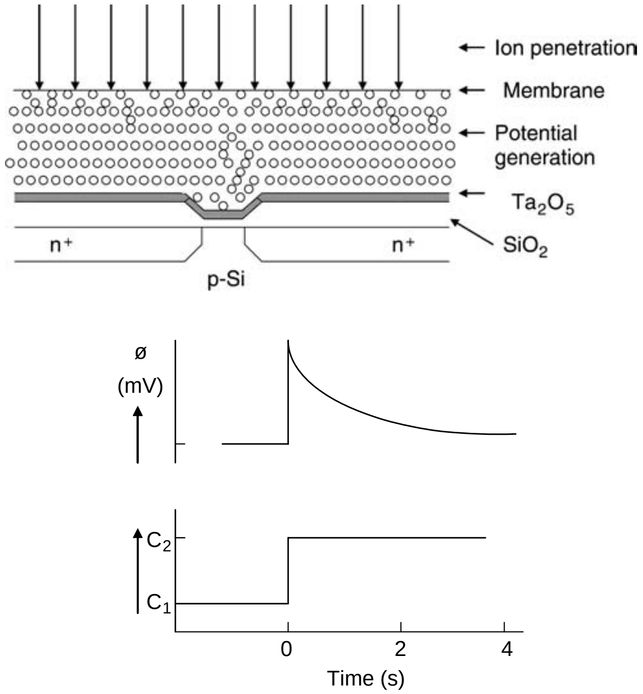


Figure 5. An ion-step arrangement: an ISFET exposed to an increased electrolyte concentration. Transient potential can bemeasured, developing from transient transport of ions across the membrane, which is caused by stepwise changes in electrolyte concentration. The ISFET response ϕ as a result of the stepwise changes in electrolyte concentration (C_1–C_2).

Indirect-Sensing IMFET

Although the ion-step method is an indirect way of measuring antigen-antibody reaction that occurs on the gate region of an ISFET, it does not involve any reagents that enhance or amplify the signal intensity. Many approaches to transduction of the antibody-antigen combining event are indirect. They are based on the ability of an enzyme label to produce electroactive substances within a short span of time. Antibody or antigen is immobilized on the gate area of pH-FET. In the competitive binding assay, the sample antigen competes with enzyme-labeled antigen for the antibody-binding sites on the membrane. The membrane is then washed, and the probe is placed in a solution containing the substrate for the enzyme. IMFETs based on the sandwich assay are applicable for measuring large antigens that are capable of binding two different antibodies. Such sensors use an antibody that binds analyte-antigen, which then binds an enzyme-labeled second antibody. After removal of the nonspecifically adsorbed label, the probe is placed into the substrate-containing solution, and the extent of the enzymatic reaction is monitored electrochemically. Gate voltage is supplied by reference electrode, such as Ag/AgCl or a Hg/Hg$_2$Cl$_2$ electrode, that is immersed in a sample solution. It is, however, difficult to make a small conventional electrode, which prevented the IMFET from being miniaturized as a whole. When a noble metal, such as platinum or gold, is used as a reference electrode, the potential between the metal electrode and sample solution fluctuates. The fluctuation makes stable measurement impossible. A method to cancel the fluctuation using a reference ISFET (REFET) is reported. A combination of two kinds of ISFET is used, one of which detects a specific substance whereas the other (REFET) does not detect it (Fig. 6). Thus, measuring the differential output between the two ISFETs can cancel the potential fluctuation in the sample solution and drift due ISFET (40–42).

Most of the indirect-sensing IMFET studies are carried out using urease-conjugated antibodies. Urea is used as a substrate. The immunosensor uses a reaction wherein urea is hydrolyzed by the urease-labeled second antibody. The

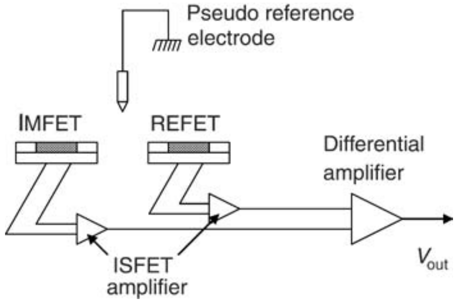


Figure 6. Differential measurement setup for an IMFET.

reaction is

$$H_2NCONH_2 + 2H_2O + H^+ \rightarrow 2NH_4^+ + HCO_3^-$$

According to the reaction, the pH value in the membrane becomes high. On the other hand, on the ISFET surface with inactive antibody membrane, the above reaction does not occur and pH remains constant. Hence, by measuring the differential output between two ISFETs, only pH changes due to urea hydrolysis are detected. In some cases, the authors used antibodies conjugated with the glucose oxidase. These sensors use oxidation of glucose by glucose oxidase. In the reaction, gluconic acid is produced and the pH value in the glucose oxidase immobilized membrane becomes low. To achieve a high sensitivity of horseradish peroxidase (HRP) detection, various substrates, either alone or in combination, are tested and the result is shown in Fig. 7.

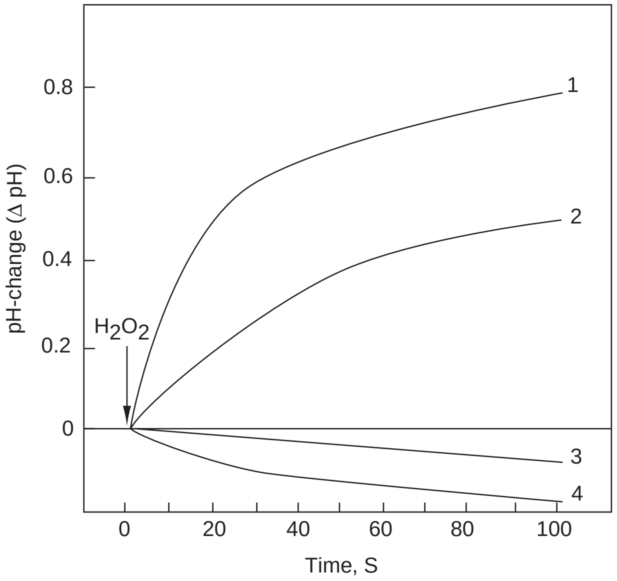


Figure 7. Typical ISFET responses for HRP (10^{-9} *M*) substrates (1) Ascorbic Acid + O-phenylenediamine (OPD); (2) OPD; (3) piodophenol+ luminol; (4) catechol.

PRACTICE

Direct-Acting IMFET

The rationale for attempting to combine the fields of immunology and electrochemistry in the design of analytical devices is that such a system should be sensitive due to the characteristics of the electrochemical detector while exhibiting the specificity inherent in the antigen-antibody reaction. The ideal situation would be to detect the binding of immunoreagents directly at an electrode, for example, by changes in surface potential, which could be truly described as an immunosensor (43,44). Much more effort has been committed to develop transducers, which rely on direct detection of antigen by the antibody immobilized on its surfaces (or vice versa). In 1975, Janata immobilized a sugar-binding protein Concanavalin A on a PVC-coated platinum electrode and studied its responses in the presence of sugar (30). The potential of the electrode with respect to an Ag/AgCl electrode changed owing to adsorption of the charged macromolecule. Although the system reported was not based on an immunochemical reaction, the finding of a potentiometric response stimulated further investigations in this field. Direct potentiometric sensing of antigen human choriogonadotropin (hCG) with an anti-hCG antibody sensitized titanium wire resulted in 5 mV shifts with respect to a saturated calomel electrode (45). The change in potential was explained by a simple charge transfer model.

In 1978, Schenck first proposed a concept of direct immunosensing by an ISFET (46,47). He suggested using FET with, on the gate region, a layer of antibody specific to a particular antigen. Replacement of electrolyte solution with another electrolyte solution-containing antigen should alter the charge of the protein surface layer due to the antigen-antibody reaction, thus affecting the charge concentration in the inversion layer of the transistor. The corresponding change in the drain current would then provide a measure of the antigenic protein concentration in the replacement solution. Many research groups have tried to realize the proposed concept of Schenck, but the results obtained are meager (48,49). Collins and Janata immobilized a PVC membrane containing cardiolipin antigen onto the gate of a previously encapsulated ChemFET (50). They demonstrated that the solution-membrane interface was somewhere between a polarized and a nonpolarized interface, based on the measured membrane exchange current density. The measured potential was therefore a mixed potential deriving out of the permeation of Na^+ and Cl^- ions into and out of the membrane. The change in potential following specific binding of antibody to the membrane was due primarily to a perturbation of the mixed potential, rather than to the adsorbed charge from the antibody itself. Therefore, the device could not be considered selective for the immunoreactive species of interest. Besides, Janata reported that it is impossible to construct an IMFET without having an ideal polarized solution-insulator interface. He proclaimed all of his earlier results as artifacts (51). In spite of these practical difficulties, Gotoh et al. (52) published results obtained with an IMFET sensitive to Human serum albumin (HSA).

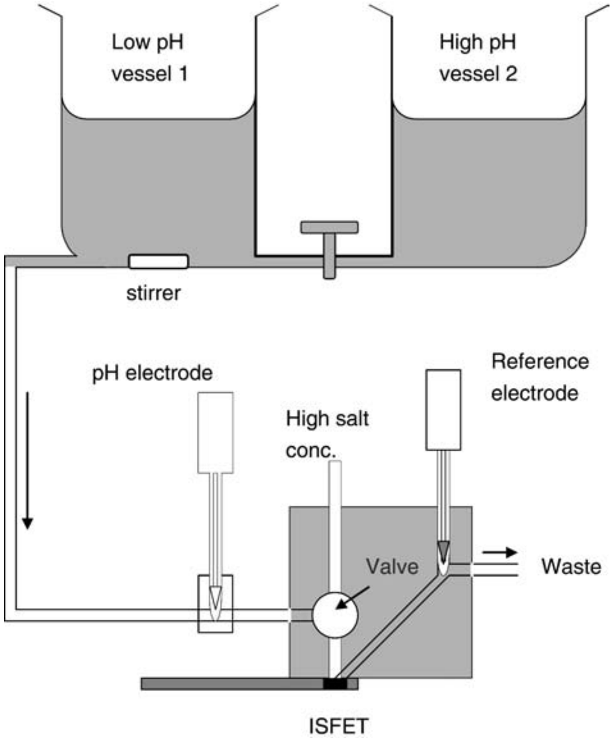


Figure 8. Outline of an Ion-step flow through system.

A 2 mV shift was detected with HSA containing polyvinylbutyral membrane deposited on an ISFET after reaction with its antibody. It appears that experimental results obtained with direct detection of protein on solid-state electrode or similar devices are, so far, limited to second-order effects. Nevertheless, a real theoretical explanation is absent. Therefore, until more experimental evidence is available, the true value of direct-acting IMFET concept remains to be established.

Schasfoort et al. (36) proposed an alternative approach to overcome the above-described difficulties of a direct detection of immunological reaction with ISFET. By stepwise changing the electrolyte concentration of the sample solution, a transient diffusion of ions through the membrane-protein layer occurs, resulting in a transient membrane potential, which can be measured by the ISFET. A flow-through system was used to carry out the experiments as schematically drawn in Fig. 8. The pH of the electrolyte can be changed by using a gradient vessel. When the solution flows under hydrodynamic pressure out of vessel 1, the pH will change through mixing with a solution of different pH from vessel 2. By opening the value for a few seconds, the ISFET can be exposed to a higher salt concentration. The step change in join concentration was completed within 50 ms. After 2 s the valve was closed and the membrane can gain equilibrate with the buffer flowing out of vessel 1. In order to exchange the electrolyte concentration rapidly, the volume between the valve and the ISFET was kept small. ISFETs with a polystyrene-agarose membrane were incubated with 10^{-5} M HSA for 3 h. The ISFET response was measured as a function of the pH, and the inversion point was determined to be pI = 3.72 ±0.05. Subsequently, the ISFETs were incubated in different concentrations of anti-HSA antibodies solution ranging from 0.06 to 64 μM. The anti-HSA antibody was able to change the inversion point of the HSA-coated membrane from 3.70 to 5.55. The above experiments clearly demonstrated that the net charge density in a protein layer deposited on an ISFET could be determined by exposing the membrane to a stepwise change in electrolyte concentration while measuring ISFET current change. The transient membrane potential observed is a result of the different mobilities of the positive and negative ions present in the protein layer. It is also observed that characteristic inversion points and slope are a function of the protein concentration and type of protein. Also isolectric points could be detected from the membrane potentials as a function of the pH. This detection of the isoelectric point of a protein complex is the basis for the development of an IMFET. An immunological reaction results in a change of the fixed-charge density in the membrane, which can be explained by a shift of the protein isoelectric point due to the immunological reaction.

The ion-step method was originally designed to measure immunoreaction via the change in charge density, which occurs in an antibody-loaded membrane, deposited on an ISFET, upon reaction with a charged antigen. The efficacy of ion-step method for the quantification of a non-charged antigen was demonstrated using progesterone as the model analyte (53). Progesterone is an uncharged molecule, hence, it cannot be detected directly by using the ion-step method. A competitive method was devised using a charged progesterone-lysozyme conjugate. To prepare the ISFETs for ion-step measurement, a membrane support was created by depositing a 1:1 mixture of polystyrene beads and agarose on the gate. The ISFETs were then cooled to 4 °C and the solvent was slowly evaporated, leaving a porous membrane with a thickness of approximately 4 μm. The ISFET was then heated to 55 °C for 1 h to immobilize the membrane onto the gate. The ISFET was placed in the flow-through system (see Fig. 8) and a monoclonal antibody specific to progesterone was incubated on the membrane (0.5 mg/ml, 4 °C for 20 h). A competitive assay method was used to detect progesterone levels, and the detection limit was approximately 10^{-8} M of progesterone in the sample solution. Recently, Besselink et al. (54) described an amino bead-covered ISFET technology for the immobilization of antibodies. HSA was immobilized onto the amino bead-coated ISFET, by covalent cross-linking method, and the anti-HSA antibodies were quantitated using the ion-step method. The antibody concentration was detected within 15 min, with yields up to 17 mV (Fig. 9).

Indirect-Sensing IMFET

The indirect-sensing IMFET concept emerged during the early 1990s in order to overcome the difficulties met with the direct-acting IMFET devices (55). Colapicchioni et al. (56) immobilized IgG using protein A onto the gate area of an ISFET. The efficacy of the IMFET was demonstrated using Human IgG and atrazine antibodies captured using protein A. As the atrazine is a small molecule (hapten), which does not induce an immunoresponse as such, it was linked to a carrier protein. Bovine Serum Albumin (BSA)

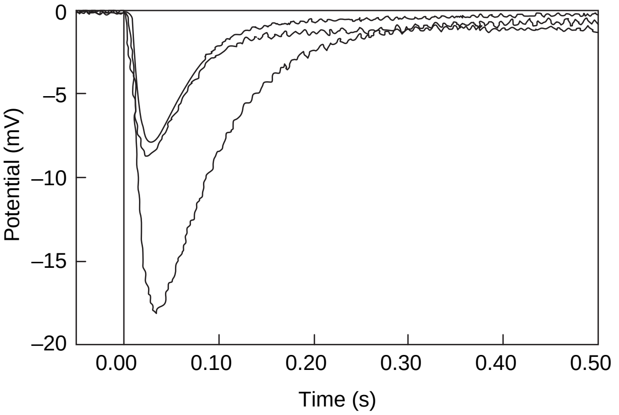


Figure 9. Ion-step responses of HSA-coated ISFET before (upper solid curve) and after incubation (for 15 min) with undiluted anti-HSA (lower solid curve) and anti-BSA (dashed curve). Ion stepping was performed at pH 4.02.

was conjugated to ametryn sulfoxide, which has structural similarity with atrazine, and the ametryn-BSA conjugate was injected into rabbit to raise antibodies. A sandwich assay format was used to detect Human IgG and a competitive assay format was used to quantitate atrazine concentration. The antigen-antibody reaction was monitored by the addition antihuman IgG-GOD conjugate and ametryn-GOD, respectively. Glucose was used as the substrate and the pH variation was detected by the ISFET. The sensitivity of the assay was 0.1 μg/ml and 1 ppb for human IgG and atrazine, respectively. An ISFET-based immunosensor was demonstrated for the detection of bacterial (*Clostridium thermocellum*) cells. The analysis included the reaction of antibacterial antibodies with cells in suspension or after covalent immobilization of cells on porous photoactivated membranes and, subsequently, the revelation of bound antibodies by the conjugate of protein A and HRP and the quantitation of enzyme activity with ISFET. The sensitivity of the sensor was within a range of 10^4–10^7 cells per ml (57). Selvanaygam et al. (23) reported ISFET-based immunosensors for the qunatitation of β-Bungarotoxin (β-BuTx), a potent presynaptic neurotoxin from the venom of *Bungarus multicinctus*. A murine monoclonal antibody (mAb 15) specific to β-BuTx was immobilized on the gate area, and the antigen-antibody reaction was monitored by the addition of urease-conjugated rabbit anti-β-BuTx antibodies. The sensor detected toxin level as low as 15.6 ng/ml. The efficacy of the sensor for the determination of β-BuTx from B. *multicinctus* venom was demonstrated in the mouse model.

An immunological *Helicobacter pylori* urease analyzer (HPUA), based on solid-phase tip coated with a monoclonal antibody specific to *H. pylori*'s urease and ISFET, was reported by Sekiguchi et al. (58). A solid-phase tip, with an inner diameter of 0.55 mm, coated with the monoclonal antibody, was incubated for 15 min at room temperature in an endoscopically collected gastric mucus sample. The activity of urease captured on the inner surface of the solid-phase tip was measured by coupling it with an ISFET in a measuring cell containing urea solution. The pH change of urea solution after 55 s of the enzymatic reaction inside the tip was measured by withdrawing 1.1 μl of solution toward the upstream of the tip, where the measuring ISFET was installed. One cycle of measurement was completed in 17.5 s, and the sensitivity of system was 0.2 m IU/ml. The calibration curve for the quantitation of urease is shown in Fig. 10. Clinical studies were carried out with 119 patients (75 males and 44 females with an average age of 51, ranging from 13 to 79) who underwent gatroduodenoscopy and judged necessary to evaluate the infection of *H. pylori* and urea breath test (UBT) was used as a gold standard. Thirty-three of the UBT positive 36 patients were positive, and 81 of UBT negative 83 patients were negative by HPUA resulting in the 92% sensitivity and 98% specificity.

An IMFET for the detection of HIV-specific antibodies based on a combination of ELISA principle and ISFET flow injection analysis setup was presented by Aberl et al. (59). The active sensing components consist of a reaction cartridge containing a carrier with the immobilized receptor layer and an ISFET sensor mounted in a flow-through cell. A flow cell was constructed using two ISFET sensors on one in a two-channel configuration (Fig. 11). The liquid

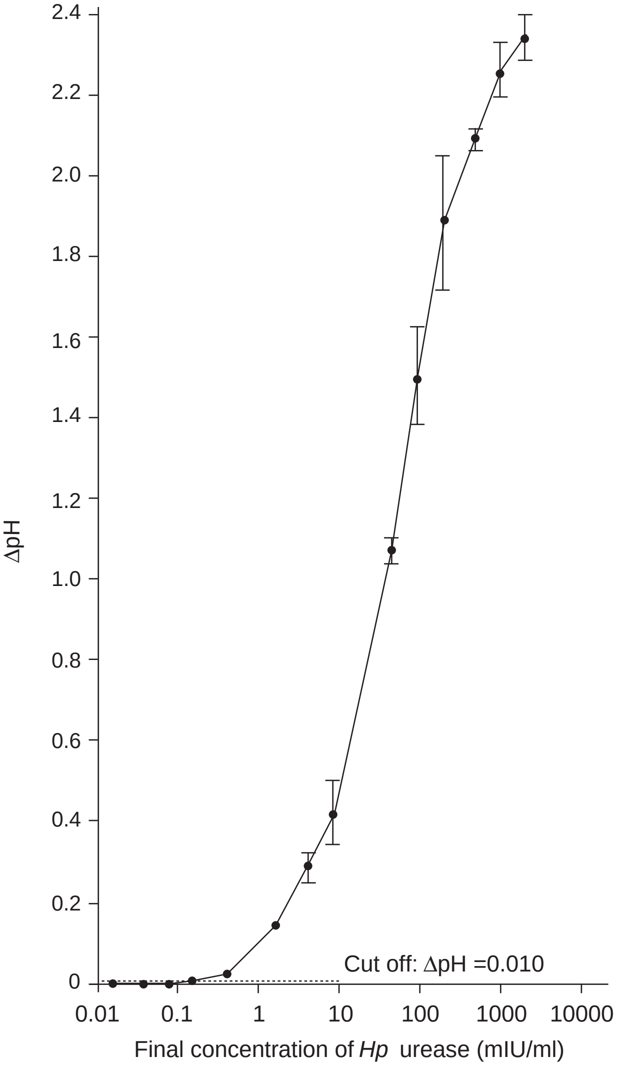


Figure 10. A standard curve for HPUA.

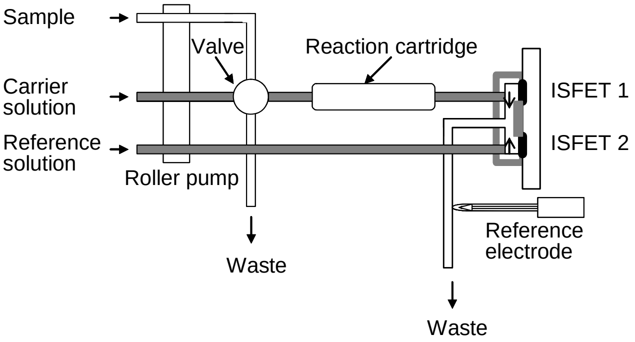


Figure 11. Diagrammatic representation of a flow injection system for indirect immunosensing.

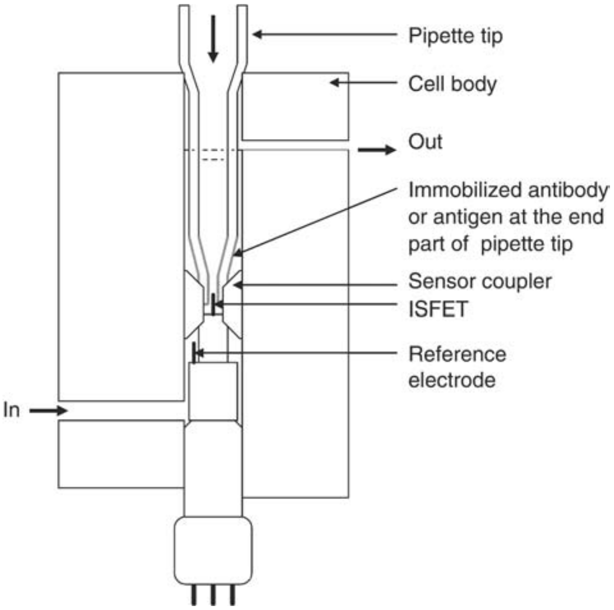


Figure 12. Cross-sectional view of a pH-measuring cell.

headspace on top of the ISFET sensors was reduced to about 1 μl, and the dead volume of the whole sensor cell was 7μl. The detection principle was realized according to the sandwich ELISA procedure using urease as a pH shifting marker enzyme. Antigen molecules (p24 or gp120) were immobilized on cellulose nitrate membranes mounted in a special flow-by cartridge or the inner surface of Borosilicate glass capillary tubing. After blocking the unspecific binding sites, the antigen was reacted with specific serum in different dilution or nonspecific serum as a negative control. In comparison with conventional ELISA, the ISFET-FIA ELISA showed a slight lower sensitivity. The antibodies were detected in a serum diluted more than 1:12,000 in ELISA, whereas the sensitivity of the ISFET– FIA ELISA was between a 1:1000 and a 1:10,000 dilution. Glass as a support material showed highly reproducible test results when compared with cellulose nitrate membrane.

Tsuruta et al. (60) reported a fully automated ISFET-based ELISA system using a pipette tip as a solid phase and urease as a detecting enzyme. The inner wall of the end part of a pipette tip was used as a solid phase, and the urease activity of the conjugate, captured after a two-step immunoreaction, was measured by coupling the pipette tip with the ISFET in a pH measuring cell (Fig. 12). A two-step sandwich assay procedure was used for the quantitation of AFP, CEA, HBsAg, and HBsAb, and a two-step competition assay was used for HBcAb, and second-antibody configuration was used for HTLV-1 Ab. After final incubation in conjugate solution, the pipette tip was washed and it was introduced into the pH measuring cell in order to couple it with ISFET. At the same time, feeding of the substrate solution was stopped, to read the pH change for 20 s. The output (source potential) of the ISFET was read and stored in the CPU during the above-mentioned 20s at 0.1 s intervals. The maximum changing rate of the source potential ($\Delta V/\Delta t$, mV/s) was calculated from these 200 data points. The total assay time was 21 min as the sum of 5, 10, 5 and 1 min for preheating of sample, First immunoreaction, Second immunoreaction, and pH measurements, respectively. The assay speed was 60 samples/h. Assay performance, such as within run CVs, between run CVs, detection limits, and correlation with the conventional ELISA kits, were satisfactory for all of six analytes. The detection limit for CEA, 0.09 μg/l was comparable to better than those reported for the most advanced chemiluminescent ELISA system (0.086 μg/l).

Polymerase chain reaction (PCR) has proven to be of great importance in clinical diagnosis (61). Usually, the PCR products have been detected by staining with ethidium bromide in qualitative methods, and fluorescent dyes in real-time quantitation. Although electrophoresis has the advantage of giving information on the molecular size of PCR products, it is not well-suited to mass screening or automation. On the other hand, real-time monitoring is well-suited for mass screening and automation but is expensive. One of the most promising methods for automatizing the detection system of PCR products is ELISA. An ISFET-based ELISA was used to quantitate PCR products (62). Double-stranded PCR products with digoxigenin and biotin at both terminals were obtained by using digoxigenin-and biotin-labeled primers. The PCR products were detected by a two-step sandwich ELISA. One μl of the solution after PCR was introduced into the end part of the solid-phase pipette tip coated with antidigoxigenin antibody. Biotin-labeled PCR products captured at the solid phase were detected with avidin-urease conjugate, and the enzyme activity was measured by the ISFET in a pH measuring cell containing urea solution. The detection limit of the system was determined using a known amount of purified PCR product labeled with digoxigenin and biotin, and it was found that 10 amol of the labeled DNA in 1 μl sample. The assay was used to detect HTLV-1 provirus gene integrated in the genome of human MT-Cell, and it was found that 100 pg of the genomic DNA was specifically detectable after 35 cycles of PCR. The apparent dynamic range for detection of MT-1 DNA was from 100 pg to 100 ng.

One of the most important targets in molecular biology is the quantitation of mRNA related to special disease by RT-PCR. The accuracy of quantitative RT-PCR has been remarkably improved by the introduction of competitive

RT-PCR, in which a synthetic RNA is used as an internal standard (63). Tsuruta et al. (64) developed a ISFET-ELISA method for the qantitatiion of mRNA in clinical samples. In this method, a fixed amount of a synthetic RNA, pRSET RNA, was added as internal standard to the solution of target RNA (IL-1β) after reverse transcription, PCR was carried out using digoxigenin-labeled sense primer and biotin-labeled antisense primer for IL-1β, and FITC-labeled sense primer and a biotin-labeled antisense primer for pRSET. The double-stranded PCR products of IL-1β and pRSET were captured by two solid-phase pipette tips, one coated with antidigoxigenin antibody and another with anti-FITC antibody, respectively, and sandwiched by an avidin-urease conjugate, whose activity was measured with ISFET. The ratio of the signal intensity for IL-1β to that for pRSET was used to quantitate the concentration of IL-1β. A calibration curve was obtained using a known amount of AW109 RNA as an external standard in place of IL-1β m RNA. It was found that 10^2–10^6 copies of IL-1β mRNA were measurable by the present method. Expression levels of IL-1β mRNA in clinical samples, such as monocytes of peripheral blood or synovial cells from patients with RA or OA, were determined.

Practical Limitations

In this section, we shall address some practical problems that have been limiting factors in the commercial application of IMFETs. The widespread use of IMFETs for applications ranging from medical diagnosis to industrial process control or environmental monitoring has not actually happened. The underlying reasons for this situation fall into two main categories, those that are inherent to the transistor, such as material, encapsulation, and reference electrode, and those problems common to its application as an immunosensor function, such as, antibody immobilization, stability, and durability. The pH sensing properties and drift behavior of the ISFET is the main limiting factor in the commercial breakthrough of ISFET. After the invention of the ISFET, initially the only gate material used was SiO_2. Although SiO_2 showed pH, sensitivity of 20 to 40 mV/pH, the thermally grown gate oxide loses its isolation property within a few hours of immersion in a solution. In order to isolate this gate oxide from the solution, another isolating layer, such as Si_3N_4, Al_2O_3, or Ta_2O_5, has to be placed on top of this gate oxide. A layer of Si_3N_4 on top of SiO_2 showed 45–50 mV/pH, and other layers, such as Al_20_3 and Ta_2O_5, showed even higher pH sensitivity, 53–57 mV/pH and 55–59 mV/pH, respectively (65). Drift rate for Si_3N_4 is reported as 1 mV/h and for Al_2O_3 and Ta_2O_5 0.1–0.2 mV/h after 1000 min of operation at pH 7.0. In most of the work on IMFETs published so far, these three gate materials, Si_3N_4, Al_2O_3, and Ta_2O_5, have been used. IMFETs are also sensitive to light and temperature (66).

The pH-sensitive ISFETs can be fabricated by means of standard MOS technology, except for the metallization step. However, after dicing the wafers into single chips, the substrate becomes exposed at the edges of the senor. Encapsulation and packaging are two final processing steps that determine reliability and durability (lifetime) of the IMFETs. In order to achieve high quality sensors, all electrical components have to be isolated from their surroundings. Several reports exist on the encapsulation and packaging of ISFET devices for pH application (21). The simplest method of isolating these sides is encapsulation with epoxy-type resins. The most important ISFET characteristics, such as stability, accuracy, and durability, also pertain to the reference electrode. One of the major hurdles in IMFETs is the lack of a solid-state reference electrode. The small IMFETs have to be combined with a conventional KCl-solution-filled reference electrode. In order to achieve miniaturized IMFET, it is important to miniaturize the reference electrode. In general, two approaches have been followed: reference FETs (REFETs), which are used in an ISFET/REFET/quasi-reference electrode setup, and miniaturized conventional reference electrodes. In the first approach, attempts have been made to cover the ISFET surface with a pH-insensitive layer or to render the surface pH insensitive by chemical modification. In the second approach, the structure of a conventional electrode (mostly Ag/AgCl type) is miniaturized partially or completely on a silicon wafer. Its potential is a function of concentration of chloride ions. They are supplied either from an internal electrolyte reservoir formed by an anisotropic etching in the silicon wafer or by adding chloride ions into the test solution.

Some of the technological factors such as pH sensitivity and drift can now be overcome with the existing technology. A hurdle peculiar to direct-acting IMFET is the need to provide a thin but fully insulating layer (membrane) between the antigen or antibody coating and the semiconductor surface. Such a membrane must be thin enough to allow a small charge redistribution occurring as a result of analyte (antigen-antibody) binding to exert a detectable change in electrical field. Conversely, it must also provide adequate insulation to prevent dissipation of the field by leakage of ions. Even assuming that the ideal insulating membrane can be developed, a further hurdle may need to be overcome. Surface charges and hydrogen binding sites of proteins cause a counter-ion shell (double-layer) and structured water shells to surround the molecules; these regions of structured charge will inevitably contribute to the electrical field affecting the FET gate. Pending these breakthroughs, the development of direct-acting IMFETs appears to be stagnant.

The immobilization methods used for immunosensors include a variety of adsorption, entrapment, cross-linking, and covalent methods. In general, a covalent immobilization method consisting of silanization step and subsequent coupling procedure via glutaraldehyde has been used to immobilize antibodies onto the gate region (67,68). However, no methodical investigation about antibody stability, storage, and lifetime exists. Reproducible regeneration of the sensing area is one of the major problems met with IMFETs that have been used for continual monitoring or repeat usage. The need for renewal of the sensing surface derives from the high affinity constants derived from the strong antigen-antibody reaction. Two different strategies have been used to achieve the renewal of the sensing surface, breakage of the antigen-antibody bond and reusing the immunologic reagent immobilized on the solid phase. A second alternative is the elimination of antigen-antibody complex from the solid support and immobilization of fresh immunologic material. Dissociation of antigen from antibody is usually carried out in low pH and high ionic strength solutions. Protein A was chemically immobilized onto the gate surface by using a polysiloxane layer of [3-(2-aminoethyl)aminopropyl]trimethoxysilane (APTES) and cross-linking agent such as glutaraldehyde. Reversibility of the linkage between Protein A and antibodies in order to restore the device for the next measurement was studied by breaking the antibody-antigen complex formed on Protein A using a variety of reagents. Glycine buffer pH 2 and 3 and $MgCl_2$ 3.5 M were found to be more effective when compared with other tested reagents due to high ionic strength (55). Selvanayagam et al. (23) studied the reusability of an ISFET sensor by removing the antibody membrane from the gate area. The regenerated devices tested were reported to function normally five times, although a considerable amount of time was required for the regeneration process. Recently, IMFET using magnetic particle and integrated to flow injection system has been described to overcome the problem of regeneration (69,70). The immunological material was immobilized on the surface of magnetic particles and were transported by a flow system, and were retained on the gate area of the ISFET by a magnetic field produced by a magnet (Fig. 13). The regeneration of immunologic materials was achieved by releasing the magnetic field, thus freeing those particles that were washed by the flow system, and new magnetic particles were injected and retained on the surface of transducer by reacting the magnetic field. A fresh and reproducible surface was thus produced, ready for the next analytical cycle.

The main barrier to the successful introduction of IMFETs for clinical testing is undoubtedly the high performance and automation level of the machines that already exist in centralized laboratories. They have been developed specifically for use with either immunoassay or clinical chemistry. Immunoassay performance is continually being optimized and assay times have been reduced over the past few years. Depending on the parameter, the assay time can be as low as 6 min and the majority of the larger machines could carry out between 100 to 200 testes per hour. IMFETs must be compared with these methods with respect to assay time, sensitivity, and cost. The need for in-built calibration has been frequently encountered in sophisticated quantitative IMFETs. Although feasible and acceptable in laboratory-based instrumentation, it remains a major problem in small disposable IMFET devices. To facilitate the increased use of IMFETs, one should look for real tests and standards that prototypes can meet, based on current diagnostic needs and perceived future development. The progress of IMFET beyond the experimental laboratory level is mainly dependent on how skillfully its development and marketing are combined with parameter selection.

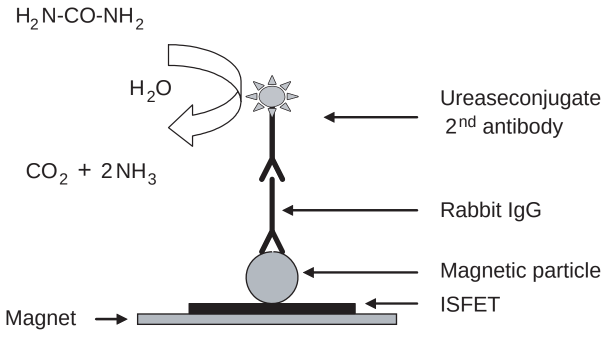


Figure 13. Schematic representation of the magnetoIMFET.

FUTURE DIRECTIONS

A key consideration in antibody immobilization to the gate area is to maintain, reproducibly, the highest possible binding activity after immobilization while conserving the amount of antibody used. However, many aspects, both fundamental and more applied, require in-depth study before IMFETs can become successful commercial device, including improved control of biomolecule immobilization and novel immobilization strategies; enhancement of biomolecule stability and retention of activity *in vitro*; and the ability to reproduce the high signal-to-noise ratios obtained in simple test solutions in "real" samples such as blood or water. The IMFETs tends to respond nonspecifically to any molecule bound to the surface; hence, it affects the measurement parameter to some extent. The specificity of analyte detection, therefore, relies entirely on achieving high ratios of specific to nonspecific binding, which, in the context of low concentrations of analyte in blood, can represent a formidable problem. Reduction of nonspecific binding is another area that will continue to be of major importance. The ability to immobilize ordered antibodies will maximize antigen binding to a given surface while reducing the availability of nonbinding site sections of the immobilized antibody, or uncovered surface areas, which can promote nonspecific interaction with other components in the sample. The potential advantages of using Fab fragments rather than more complex intact antibodies (such as IgG) could be explored.

IMFETs, similar to immunoassays, involve multistep procedures, including washing steps, incubation periods, and quite complex signal generation protocols. It is likely that research efforts into a novel signal amplification system, without the normally associated complications of multi-reagents or multistep protocol will be of increasing importance. The irreversibility of antigen-antibody interaction presents a major challenge in designing IMFETs for continual monitoring or repeated usage. Treatment of an antibody-antigen complex with a mildly denaturing medium for a short time interval has shown some promise in regenerating sensor surfaces. Development of enhanced denaturation conditions, which optimize dissociation of antigen while minimizing irreversible loss of antibody structural integrity, may be possible in the near future. The use of catalytic antibodies in immunosensors has been proposed. The ability of catalytic antibodies to catalyze the hydrolysis of phenyl acetate with the formation of acetic acid allows integration of pH-sensitive microelectrodes to give a potentiometric immunosensing system (71). The advantage of catalytic antibodies over normal antibodies is that reversibility of response is readily achieved, because bound antigen reacts to form a product with a low affinity for the antibody, resulting in dissociation. As the binding event is followed immediately by a catalytic reaction and release of the reaction products, the molecular recognition site is regenerated with each molecular reaction; as a consequence, catalytic antibodies can be used to create reversible IMFETs for continuous monitoring of analyte concentrations. Improvements in sensitivity and cross-reactivity are likely to be achieved as a result of the increasing interest in this field of research.

CONCLUSION

Although the ISFET concept has existed for over 30 years, its practical applications such as the IMFETs are still emerging very slowly. The relatively slow rate of progress of IMFET technology from inception to fully functional commercial devices for these applications is a reflection of technology-related and market factors. In general, two approaches have been followed in the past to realize IMFETs. In the first approach, antigen-antibody reaction on an immonbilized membrane was monitored without any addition of labels. The second approach takes the advantage of an enzyme label to indirectly monitoring the antigen-antibody reaction using pH-sensitive FET. The development of IMFETs that directly detect antigen-antibody reaction is extremely challenging; only a few examples exist, the majority of which are without valid theoretical explanation. Although it shows enormous promise in the early stages of development, an effective, reliable, and analyte-selective direct-acting IMFET sensor is yet to be constructed. The ion-step method represents a novel measurement concept for potentiometric detection and quantification of an adsorbed antigen or antibody molecule in which modified ISFETs are used. Many approaches to transduction of the antibody-antigen combining event are indirect, necessarily involving the use of reagents admixed with analyte, and therefore cannot be seen as routes to development of “True” IMFETs. Nevertheless, such reagent-dependent, indirect-sensing IMFETs may offer real commercial advantages over the current generation direct-acting IMFET readout technologies. The clinical diagnostic field offers real opportunities for the exploitation of IMFET, but because it is a highly competitive and well-established market, those who wish to introduce new products must carefully target their market niche. IMFETs will have to compete with such technology on the basis of factors such as cost, ease of use, sensitivity, operational stability, robustness, and shelf-life.

BIBLIOGRAPHY

1. Catty D. editor, Antibodies Volume 1. A Practical Approach. Washington, DC: IRL; 1988.
2. Mayforth RD. Designing Antibodies. New York: Academic Press; 1993.
3. Johnstone AP, Turner MW. Immunochemistry 1. A Practical Approach. New York: IRL; 1997.
4. Bluestein BI, Walczak IM, Chen SY. Fiber optic evanescent wave immunosensors for medical diagnostics. Trends Biotechnol 1990;8:161–168.
5. Gosling JP, A decade of development in immunoassay methodology. Clin Chem 1990;36:1408–1427.
6. Kemeny DM, A Practical Guide to ELISA. New York: Pergamon Press; 1991.
7. Turner APF, Karube I, Wilson GS. editors. Biosensors: Fundamentals and Applications. New York: Oxford University Press; 1987.
8. Buerk DG, Biosensors: Theory and Applications. Lancester, PA: Technomic; 1993.
9. Eggins BR. Biosensors: An Introduction. New York: Wiley; 1996.
10. Aizawa M. Principles and applications of electrochemical and optical biosensors. Anal Chim Acta 1991;250:249–256.
11. North JR. Immunosensors: Antibody-based biosensors. Trends Biotechnol 1985;3:180–186.
12. Aizawa M. Immunosensors for clinical analysis. Adv Clin Chem 1994;31:247–275.
13. Morgan LC, Newman DJ, Price CP. Immunosensors: Technology and opportunities in laboratory medicine. Clin Chem 1996;42:193–209.
14. Ghindilis AL, Atanasov P, Wilkins M, Wilkins E. Immunosensors: Electrochemical sensing and other engineering approaches. Biosens Bioelecton 1998;13:113–131.
15. Bergveld P. Future applications of ISFETs. Sens Actuators B 1991;4:125–133.
16. Kuriyama T, Kimura J. An ISFET biosensor. In: Wise DL, editor, Applied Biosensors. Boston, MA: Butterworth; 1989.
17. Kimura J, Kuriyama T. FET biosensors. J Biotechnol 1990;15:239–254.
18. Van der Schoot BH, Bergveld P. ISFET based enzyme sensors. Biosensors 1987/88;3:161–186.
19. Yuquing M, Jianguo G, Jianrong C. Ion-sensitive field effect transducer-based biosensors. Biotechnol Advances 2003; 21:527–534.
20. Janata J, Moss SD. Chemically sensitive field-effect transistors. Biomed Eng 1976;11:241–245.
21. Bergveld P, Sibbald A. Comprehensive Analytical Chemistry, XXIII: Analytical and Biomedical Applications of Ion-Selective Field-Effect Transistors. New York: Elsevier; 1988.
22. Bergveld P. Thirty years of ISFETOLOGY. What happened in the past 30 years and what may happen in the next 30 years. Sens Actuators B 2003;88:1–20.
23. Selvanayagam ZE, Neuzil P, Gopalakrishnakone P, Sridhar U, Singh M, Ho LC. An ISFET-based immunosensor for the detection of β-Bungaraotoxin. Biosens Bioelectron 2003; 10:405–414.
24. Janata J, Huber RJ. Chemically sensitive field effect transistors. In: Freiser H, editor. Ion-Sensitive Electrodes in Analytical Chemistry, vol. II, New York: Plenum Press; 1980.
25. Janata J, Blackburn GF. Immunochemical potentiometric sensors. Ann N Y Acad Sci 1984;428:286–292.
26. Blackburn GF. Chemically sensitive Field effect transistors. In: Turner APF, Karube I, Wilson S. editors. Biosensors: Fundamentals and Applications. Oxford: Oxford Science Publications; 1987.
27. Bockris JOM, Reddy AKN. Modern Electrochemistry, vol. 2. New York: Plenum Press; 1970.
28. Eisen HN. Immunology. An Introduction to Molecular and Cellular Principles of the Immune Response. New York: Harper and Row; 1974.
29. Bergveld P. A critical evaluation of direct electrical protein detection methods. Biosens Bioelecton 1991;6:55–72.
30. Janata J. An immunoelectrode. J Am Chem Soc 1975;97:2914–2916.

31. Janata J, Janata J. Novel protein-immobilizing hydrophobic polymeric membrane, process for producing same and apparatus employing same. U.S. Patent 3,966,580, 1976.
32. Schasfoort RBM, Bergveld P, Kooyman RPH, Greve J. Possibilities and limitations of direct detection of protein changes by means of immunological field-effect transistor. Anal Chim Acta 1990;238:323–329.
33. Moore WJ. Physical Chemistry. London: Longman; 1976.
34. Davies JT, Rideal EK. Interfacial Phenomena. London: Academic Press; 1963.
35. Schasfoort RBM, Bergveld P, Bomer J, Kooyman RPH, Greve J. Modulation of the ISFET response by an immunological reaction. Sens Actuators 1989;17:531–535.
36. Schasfoort RBM, Kooyman RPH, Bergveld P, Greve J. A new approach to ImmunoFET operation. Biosens Bioelecton 1990; 5:103–124.
37. Schasfoort RBM, Eijsma B. The ions-step method applied to disposable membranes for the development of ion-responding immunosensors. Sens Actuators 1992;6:308–311.
38. Eijkel JCT, Olthuis W, Bergveld P. An ISFET-based dipstick device for protein detection using the ion-step method. Biosens Bioelectron 1997;12:991–1001.
39. Kharitonov AB, Wasserman J, Katz E, Willner I. The use of impedance spectroscopy for the characterization of protein-modified ISFET devices: Application of the method for the analysis of biorecognition processes. J Phys Chem 2001; B-105:4205–4213.
40. Sergeyeva TA, Soldatkin AP, Rachkov AE, Tereschenko MI, Piletsky SA, Elskaya AV. β-Lactamase label-based potentiometric biosensor for α-2 interferon detection. Anal Chim Acta 1999;390:73–81.
41. Yacoub-George E, Wolf H, Koch S, Woias P. A miniaturized ISFET-ELISA system with a pre-treated fused silica capillary as reaction cartridge. Sens Actuators B 1996;34:429–434.
42. Starodub NF, Dzantiev BB, Starodub VN, Zherdev AV. Immunosensor for the determination of the herbicide simazine based on an ion-selective field-effect transistor. Anal Chim Acta 2000;424:37–43.
43. Wang J. Electroanalytical Techniques in Clinical Chemistry and Laboratory Medicine, New York: VCH; 1988.
44. Luppa PB, Sokoll LJ, Chan DW. Immunosensors — principles and applications to clinical chemistry. Clin Chim Acta 2001;314:1–26.
45. Yamamoto N, Nagasawa Y, Sawai M, Sudo T, Tsubomura H. Potentiometric investigations of antigen-antibody and enzyme-enzyme inhibitor reactions using chemically modified metal electrodes. J Immunol Methods 1978;22:309–317.
46. Schenck JF. Technical difficulties remaining to the application of ISFET devices. In: Cheung PW, editor. Theory, Design and Biomedical Applications of Solid State Chemical Sensors. New York: CRC; 1978.
47. Schenck JF. Field effect transistor for detection of biological reactions. U.S. Patent 4,238, 757, 1980.
48. Schoning MJ, Poghossiana A. Recent advances in biologically sensitive field-effect transistors (BioFETs). Analyst 2002; 127:1137–1151.
49. Janata J, Thirty years of CHEMFETs — a personal view. Electroanalysis 2004;16:1831–1835.
50. Collins S, Janata J. A critical evaluation of the mechanism of potential response of antigen polymer membrane to the corresponding antiserum. Anal Chim Acta 1982;136:93–99.
51. Janata J. Proceedings of the 2nd International Meeting on Chemical Sensors; 1986:25–31.
52. Gotoh M, Tamiya E, Karube I. Micro-FET biosensors using polyvinylbutyral membrane. J Membrane Sci 1989;41:291–303.
53. Schasfoort RBM, Keldermans CEJM, Kooyman RPH, Bergveld P, Greve J, Competitive immunological detection of progesterone by means of the ion-step induced response of an immunoFET. Sens Actuators 1990;BI:368–372.
54. Besselink GAJ, Schasfoort RBM, Bergveld P. Modification of ISFETs with a monolayer of latex beads for specific detection of proteins. Biosens Bioelecton 2003;18:1109–1114.
55. Toshihide K. Electrochemical sensors in immunological measurement. E.U. Patent 0,328,380, 1998.
56. Colapicchioni C, Barbaro A, Porcelli F, Giannini I. Immunoenzymatic assay using CHEMFET devices. Sens Actuators B 1991;6:186–191.
57. Akimeno VK, Khomutov SM, Obraztsova AY, Vishnivetskii SA, Chuvilskaya NA, Laurinavichus KS, Reshetilov AN. A rapid method for detection of Clostridium thermocellum by field-effect transistor immunodetection. J Microbiol Methods 1996;24:203–209.
58. Sekiguchi T, Nakamura M, Kato M, Nishikawa K, Hokari K, Sugiyama T, Asaka M. Immunological Helicobacter pylori urease analyzer based on ion-sensitive field effect transistor. Sens Actuators 2000;B67:265–269.
59. Aberl F, Modrow S, Wolf H, Koch S, Woias P. An ISFET-based FIA system for immunosensing. Sens Actuators 1992;B6:186–191.
60. Tsuruta H, Yamada H, Motoyashiki Y, Oka K, Okada C, Nakamura M. An automated ELISA system using a pipette tip as a solid phase and a pH-sensitive field effect transistor as a detector. J Immunol Methods 1995;183:221–229.
61. Saiki R, Scharf S, Faloona F, Mullis K, Horn G, Erlich H, Arnheim N. Enzymatic amplification beta-globulin genomic sequences and restriction analysis for diagnosis of sickle cell anemia. Science 1985;230:1350–1354.
62. Tsuruta H, Matsui S, Hatanaka T, Namba T, Miyamoto K, Nakamura M. Detection of the products of polymerase chain reaction by an ELISA system based on an ion-sensitive field effect transistor. J Immunol Methods 1994;176:45–52.
63. Wang AM, Doyle MV, Mark DF. Quantitation of mRNA by the polymerase chain reaction. Proc Natl Acd Sci USA 1989; 86:9717–9721.
64. Tsuruta H, Matsui S, Oka K, Namba T, Shinngu M, Nakamura M. Quantitation of IL-1-beta mRNA by a method of RT-PCR and an ELISA based on ion-sensitive field effect transistor. J Immunol Methods 1995;180:259–264.
65. Gopel W, Hesse J, Zemel JN editorss. Sensors: A Comprehensive Survey, Chemical and Biochemical Sensors, Part I. Verlagsgesellschaft Mbh: VCH; 1991.
66. Neuzil P. ISFET integrated sensor technology. Sens Actuators 1995;B24-25:232–235.
67. Filippo P, Antonio NC. Sensor with antigen chemically bonded to a semiconductor device. E.U. Patent 0,395,137, 1990.
68. Starodub NF, Samodumova IM, Starodub VN. Usage of organosilanes for integration of enzymes and immunocomponents with electrochemical and optical transducers. Sens Actuators 1995;B24-25:173–176.
69. Santandreu M, Sole S, Fabregas E, Alegret S. Development of electrochemical immunosensing systems with renewable surfaces. Biosens Bioelectron 1998;13:7–17.
70. Sole S, Alegret S, Cespedes F, Fabregas E. Flow injection immunoanalysis based on a magnetoimmunosensor system. Anal Chem 1998;70:1462–1467.
71. Blackburn GF, Talley DB, Booth PM, Durfor CN, Martin MT, Napper AD, Rees AR. Potentiometric biosensor employing catalytic antibodies as the molecular recognition element. Anal Chem 1990;62:2211–2216.

See also Ion-sensitive field effect transistors.

IMMUNOTHERAPY

Qiao Li
University of Michigan
Ann Arbor, Michigan

INTRODUCTION

Immunotherapy of cancer, infectious disease, and autoimmune disease has opened a new area for disease management. This approach has developed very fast lately due to the advances and involvements of modern technology in molecular biology, cell biology, immunology, biochemistry, and bioengineering. Adoptive T cell immunotherapy of cancer involves passive administration of lymphoid cells from one host to another, or back to itself in order to transfer tumor immunity for cancer treatment. It was first realized >20 years ago that adoptive immunotherapy may be feasible to treat human malignancies. However, the early form of this practice was quite simple. It could be as easy as a straightforward blood cell transfer. The apparent inefficiency of antitumor immune responses, and the failure to successfully combat the disease laid the foundation for current concepts of immunotherapy. It did not take too long before it was realized that boosting the antitumor immune response by deliberate vaccination could increase the potential benefits of immune cell-based therapies. In addition, activation of lymphoid cells with monoclonal antibodies (mAb) toward the molecules involved in T cell signaling pathways has resulted in therapeutic effector T cells. The use of immune adjuvants coadministrated with the cell infusion has enhanced the antitumor efficacy of the transferred cells and has made adoptive cellular immunotherapy a promising strategy for cancer treatment. Studies on the trafficking of adoptively transferred cells *in vivo* as well as the identification and characterization of T cell subsets responsible for antitumor reactivity have provided valuable insights toward the development of novel immunotherapeutic strategies. The adoptive immunotherapy of established tumors with the transfer of tumor-reactive lymphoid cells has now been shown to be highly effective against significant tumor burdens both in animal models and in clinical trials. This is, at least in part, due to recent developments in this area, such as treating cancer in special settings (e.g., in lymphopenic hosts induced by prior conditioning); redirecting the effector cells to tumor through genetic engineered chimerical T cell receptors (TCRs) or by transferred tumor antigen-specific TCRs; and the use of these strategies in combination. This article intends to review the above developments that have made adoptive T cell immunotherapy an attractive alternative for cancer treatment.

INDUCTION OF TUMOR-REACTIVE PRE-EFFECTOR T CELLS *IN VIVO*

Successful induction of tumor-reactive "pre-effector“ cells in a tumor-bearing host represents the first step toward the conduct of an effective adoptive T cell immunotherapy of cancer. This procedure provides a source of "pre-effector" cells for subsequent T cell activation and expansion *in vitro* to generate large numbers of "effector" cells to be infused back to the tumor-bearing host or cancer patient for therapy. Due to the relative lack of immunogenicity and potential immunosuppressive mechanisms of human malignancies, application of tumor T cell therapy in the clinical setting has been hampered for a long time by difficulties to reliably isolate tumor-sensitized lymphoid cells from the cancer-bearing host. Nevertheless, recent observations in animal studies and clinic trials have led to the development of strategies to induce T cell sensitization *in vivo*.

Peripheral blood lymphocytes (PBL) represents a convenient source of pre-effector cells. However, in most cases, particularly in the case of solid tumors, the frequency of tumor-specific pre-effector cells in PBL is extremely low, generally far below what is observed in response to viral infection. Experimental studies and clinical experience with adoptive immunotherapy have demonstrated that tumor-draining lymph node (TDLN) cells are potentially effective antitumor reagents. Chang et al. was the first to evaluate vaccine-primed LN (VPLN) as a source of lymphoid cells that could be secondarily sensitized by *in vitro* methods to generate effector cells capable of mediating regression of established tumors upon adoptive transfer in clinical trials (1–3). These trials included subjects with metastatic melanoma, renal cell cancer, and head and neck squamous cell cancers, and have resulted in prolonged, durable, complete responses.

In murine models, it has been observed that TDLN harbor lymphoid cells that are functionally capable of mediating rejection of immunogenic tumors in adoptive transfer after *in vitro* activation (4,5). However, both tumor-infiltrating lymphocytes (TIL) and TDLN cells were found to be incapable of mediating the regression of poorly immunogenic tumors such as the B16–BL6 melanoma, a highly invasive tumor of spontaneous origin. It was then discovered that the subcutaneous inoculation of B16–BL6 tumor cells admixed with the bacterial adjuvant, *Corynebacterium parvum*, resulted in reactive TDLN cells that differentiated into therapeutic effector T cells upon activation *in vitro* (6,7). Upon adoptive transfer, these LN cells successfully mediated the regression of established tumors. In addition to the ability to mediate regression of experimentally induced pulmonary metastases, these activated cells were effective in the treatment of spontaneous visceral metastases originating from a primary tumor, a condition that more closely approximates human malignancy. These studies thus demonstrated that vaccination of animals with irradiated tumor cells admixed with a bacterial adjuvant was capable of inducing tumor-reactive T cells in the draining LN.

We have applied these methods to generate vaccine-primed LN in patients with advanced melanoma and renal cell cancer (RCC) for therapy (3,8). Patients with RCC or melanoma received intradermal inoculation of irradiated autologous tumor cells admixed with *Bacillus Calmette–Guerin* (BCG) as a vaccine. Seven to ten days later, draining LN were removed for *in vitro* activation and expansion. Activated LN cells were then administrated intravenously with the concomitant administration of IL-2 for immunotherapy with defined success (3).

Studies demonstrated that tumor cells genetically modified with immunostimulatory genes are capable of sensitizing T cells. Transfection of cytokine genes into murine tumor cells have resulted in reduced tumorigenicity following inoculation of the modified tumor cells into animals (9). In these studies, animals that rejected the inoculum of modified tumor cells also rejected a subsequent challenge of unmodified parental tumor cells, thus demonstrating the development of tumor immunity. We performed a clinical study of patients with melanoma to evaluate the immunobiological effects of GM–CSF transduced autologous tumor cells given as a vaccine to prime draining lymph nodes (10). There was an increased infiltration of dendritic cells (DCs) in the GM–CSF-secreting vaccine sites compared with the wild type (WT) vaccine sites. This resulted in a greater number of cells harvested from the GM–CSF–VPLNs compared with the WT–VPLNs. Patients received adoptively transferred GM–CSF–VPLN cells secondarily activated and expanded *in vitro*. A complete clinical response was observed in one of five patients. This work documented measurable immunobiologic differences of GM–CSF-transduced tumor cells given as a vaccine compared with WT tumor cells.

Collectively, these observations suggested that TDLN or VPLN cells may represent an ideal source of tumor-reactive T cells. They also established the rationale for developing tumor vaccines utilizing autologous tumors admixed with bacterial adjuvant or genetically modified with cytokine genes, which may prove useful in facilitating the generation of immune T cells for adoptive immunotherapy.

ACTIVATION AND POLARIZATION OF EFFECTOR T CELLS *IN VITRO*

A major challenge in T cell immunotherapy of cancer is how to activate and expand the relatively low numbers of tumor-specific T cells obtained from the tumor-bearing host. Previous studies demonstrated that freshly isolated TDLN cells had defects in TCR-mediated signal transduction and were not immediately competent in adoptive transfer models (11,12). It has therefore become a critical prerequisite in adoptive immunotherapy to expand the pre-effector cells into large numbers of effector cells while augmenting their antitumor reactivity.

In vitro T cell activation using monoclonal antibodies in the absence of antigen takes advantage of common signal transduction pathways that are ubiquitous to T cells. This principle has been used to expand tumor-primed T cells contained within TDLN or VPLN. The initial efforts involved the use of anti-CD3 mAb as a surrogate antigen to activate tumor-primed lymphoid cells, followed by expansion in IL-2 (12). This approach resulted primarily in the generation of CD8$^+$ effector cells that mediated tumor regression *in vivo*. Subsequent clinical studies utilizing this method to activate VPLN cells demonstrated that this cellular therapy can result in achieving durable tumor responses in subjects with advanced cancer (1,3). We have extended these investigations in animal models and with human samples by examining other mAbs that deliver costimulatory signals in concert with anti-CD3 to activate tumor-primed lymphoid cells. These other antibodies have involved anti-CD28 and anti-CD137 (13–16). The results of these investigations have indicated that costimulation can increase the proliferation of tumor-primed lymphoid cells and their ability to mediate tumor regression *in vivo*.

Several important principles in animal models that are relevant for the treatment of human malignancies have been identified. For example, the *in vitro* cytokine profiles released by effector T cells when cocultured with tumor cells are found to be predictive of their ability to mediate tumor regression *in vivo*. Effector cells that mediate a type 1 (i.e., IFNγ) and GM–CSF response to tumor antigen are capable of eradicating tumor upon adoptive transfer. In contrast, cells that demonstrate a type 2 profile (i.e., IL-10, IL-4) appear suppressive, and do not mediate tumor regression (16,17). We have determined the importance of IFNγ in mediating tumor regression both in animal studies (16) and in clinical trials (3). In a phase II adoptive cellular trial in patients with advanced renal cell cancer, we demonstrated that IFNγ secretion and the IFNγ: IL-10 ratio of cytokine released by effector T cells in response to tumor antigen was associated with clinical outcomes. Specifically, activated T cells that have an increased IFNγ:IL-10 ratio correlated with tumor response (3). Although effector T cells can be generated through antibody activation to mediate tumor regression in animal models, clinical responses in adoptive immunotherapy have been confined to a minority of patients. One potential reason for these limited responses is that antibody-activation procedures generally stimulate T cells broadly without discriminating between type1 and type 2 cells, presumably due to the polyclonal expansion characteristics of antibodies directed to the TCR common chain, for example, CD3ε of the TCR/CD3 complex or CD28. As a result, both type 1 cytokines, such as IL-2, IFNγ, and type 2 cytokines, for example, IL-4, IL-5, and IL-10, are modulated (13,18). Therefore, alternative protocols need to be defined that will preferentially stimulate the type 1 cytokine profile to generate more potent tumor-reactive T cells for cancer immunotherapy. Toward this end, various *in vitro* strategies have been investigated utilizing additional signaling stimuli to promote Th1/Tc1 cell proliferation and antitumor reactivity (19,20). We reported that costimulation of TDLN cells through newly induced 4-1BB and CD3/CD28 signaling can significantly increase antitumor reactivity by shifting T cell responses toward a type 1 cytokine pattern, while concomitantly decreasing type 2 response (16). Using the proinflammatory cytokines, we recently reported that IL-12 and IL-18 can be used to generate potent CD4$^+$ and CD8$^+$ antitumor effector cells by synergistically polarizing antibody-activated TDLN cells toward a Th1 and Tc1 phenotype, and that the polarization effect was NF-κB dependent (21).

The recognition and use of cell polarization strategies during and /or post antibody activation of T cells represents another significant change and addition to the traditional practice of adoptive therapy. While adoptive immunotherapy of cancer requires large numbers of therapeutic T cells for transfer into cancer patients, the phenotype and cytokine profile of these cells are crucial in determining the outcomes of the therapy. The use of polarizing reagents to modulate T cell function toward the type 1 response provides a rational strategy to enhance the efficacy of cellular therapy.

USE OF IMMUNE ADJUVANT IN CONCERT WITH T CELL ADMINISTRATION

In the course of adoptive immunotherapy of cancer, administration of T cell growth factors accompanying T cell transfer may promote T cell activation, proliferation, and tumor killing, and therefore augment clinical outcomes for the therapy. These growth factors, as well as other immune modulatory reagents used in concert with T cell transfer, function as immune adjuvants in eliciting antitumor activities *in vivo*. The most useful adjuvant to T cell transfer to date has been the exogenous administration of IL-2 (1–3,22,23).

Nearly 20 years ago, Rosenberg and colleagues performed a pilot protocol to investigate the feasibility and practicality of immunotherapy of patients with advanced cancer using TIL and recombinant IL-2 (22). The study represents an initial attempt to use TIL plus IL-2 administration with enhanced tumoricidal capacity in the adoptive immunotherapy of human malignancies. Twelve patients with melanoma, renal cell carcinoma, breast carcinoma, or colon carcinoma were treated with varying doses and combinations of TIL, IL-2, and cyclophosphamide. Three partial responses (PR) to therapy were observed. No toxic effects were directly attributable to TIL infusions. However, the toxicities of therapy were similar to those ascribed to IL-2. Indeed, the use of IL-2 has resulted in significant morbidity associated with cellular therapies (3,24). Moreover, a few recent studies showed that IL-2 may negatively regulate effector cells through activation-induced cell death (23,25), expanding the frequency of $CD4^+CD25^+$ T cells, or cause cell redistribution secondary to Ag-induced cell death (26,27). These studies suggest that novel reagents need to be identified to serve as alternative immune adjuvants for adoptive T cell therapy.

In a recent study (25), failed adoptive T cell therapy could be reversed with low dose IL-15 administration, but not IL-2. A related T cell growth factor, IL-15, protected T cells against activation-induced cell death and promoted homeostatic maintenance of memory T cells and, therefore, may be advantageous to T cell-based cancer treatment. Similarly, the role of IL-15 in early activation of memory $CD8^+$ CTLs has been described (28). In this study, memory $CD8^+$ T cells expressing OVA-specific TCR were transferred into IL-15-transgenic (Tg) mice, IL-15 knockout (KO) mice, or control C57BL/6 mice followed by challenge with recombinant Listeria monocytogenes expressing OVA (rLM-OVA). *In vivo* CTL activities were significantly higher in the IL-15 Tg mice, but lower in the IL-15 KO mice than those in control mice at the early stage after challenge with rLM-OVA. *In vivo* administration of rIL-15 conferred robust protection against reinfection via activation of the memory $CD8^+$ T cells. In addition, IL-27 is a novel IL-12 family member that plays a role in the early regulation of Th1 initiation and synergizes with IL-12 in IFNγ production (29). Mechanistic studies revealed that although a comparable proliferative response to IL-27 was observed between STAT1-deficient and wild-type $CD4^+$ T cells, synergistic IFNγ production by IL-27 and IL-12 was impaired in STAT1-deficient $CD4^+$ T cells. IL-27 also augmented the expression of MHC class I on $CD4^+$ T cells in a STAT1-dependent manner (29).

Although the *in vivo* administration of proinflammatory cytokines has demonstrated antitumor efficacy, their potent antitumor activity is often achieved at the expense of unacceptable toxicity. For example, IL-12 and IL-18 administration was found to be associated with lethal organ damages, attributed in part to extremely high levels of host-induced IFNγ production (30). It is anticipated that administration of low doses of proinflammatory cytokines in the context of passively transferred TDLN cells will lead to increased therapeutic efficacy. To this end, the adjuvant effect of low dose cytokine administration over a long period of time can be compared with that of a high dose over a short period of time. These experiments should help to determine if prolonged administration of low dose cytokines can enhance the therapeutic efficacy by improving trafficking, survival, and proliferation of the adoptively transferred T cells.

While toxicity of traditionally used IL-2 limits its clinical utility at high doses, use of novel cytokines at tolerable low doses in conjunction with cellular therapy may provide alternative strategies that are less toxic. If the newly identified proinflammatory cytokines, such as IL-15 and IL-27 prove to be useful adjuvants to T cell therapy, they may result in more effective antitumor responses with reduced morbidity.

TRAFFICKING AND PROLIFERATION OF EFFECTOR T CELLS AFTER ADOPTIVE TRANSFER

Adoptive T cell therapy has been used for treatment of viral and malignant diseases with encouraging results. However, little is known about the fate and trafficking of the transferred effector cells. A study performed at NCI assessed the trafficking of gp100-specific pmel-1 cells to large, vascularized tumors that express or do not express the target Ag (31). It was found that approximately equal numbers of pmel-1 T cells infiltrated the Ag-positive and -negative tumors. Massive infiltration and proliferation of activated antitumor pmel-1 cells were observed in a variety of peripheral tissues, including lymph nodes, liver, spleen, and lungs, but not peripheral blood. However, T cell function, as measured by production of IFNγ, release of perforin, and activation of caspase-3 in target cells, was confined to Ag-expressing tumor. It was thus concluded that adoptively transferred $CD8^+$ T cells traffic indiscriminately and ubiquitously while mediating specific tumor destruction.

We recently characterized the infiltration of adoptively transferred TDLN cells in the host bearing pulmonary metastases (21). The TDLN cells were activated with anti-CD3/anti-CD28 followed by cell culture in IL-12 + IL-18 before transfer into tumor-bearing host. The TDLN cells were labeled with CFSE immediately before infusion. Immunohistochemical evaluation of adoptively transferred TDLN cells accumulating in pulmonary tumor nodules was performed. Infused TDLN cells were observed to (1) attach to venules, (2) mix with host leukocytes in perivenular collections, and (3) infiltrate tumor nodules. Active migration of infused cells into pulmonary tumor nodules was found to be correlated with significant tumor regression. This corroborates a previous report by Plautz et al. showing that infused TDLN cells must infiltrate pulmonary nodules to suppress tumor growth (32).

Several other reports support the hypothesis that efficient tumor regression needs the *in situ* accumulation of transferred effector cells. Another study conducted at the University of Michigan demonstrated that the infused cells must accumulate in metastatic lesions to suppress tumor growth, and that the process is dynamic (33). In studies treating murine lung metastases with adoptively transferred TDLN cells, the TDLN donor cells were initially confined to alveolar capillaries with no movement into metastases after infusion. However, within 4 h, TDLN cells began migrating across pulmonary postcapillary venules and first appeared within metastases. After 24 h, most donor cells in the lung were associated with tumor nodules. Donor cell proliferation both within the lung and in the lymphoid organs was detected. Importantly, T cells that had proliferated in the lymphoid organs trafficked back to the tumor-bearing lungs, accounting for $\sim 50\%$ of the donor cells recovered from these sites. These studies demonstrate that adoptively transferred TDLN cells migrate directly into tumor-bearing organs and seed the recirculating pool of lymphocytes after infusion. Cells that have differentiated in lymphoid organs eventually migrate into the tumor site. Additionally, *in vitro*-generated Melan-A-specific CTLs were found to survive intact *in vivo* for several weeks and localize preferentially to tumor (34). Over all, these studies suggest that methods to improve trafficking and recruitment of donor T cells to the tumor may improve therapeutic efficacy of cellular therapy.

The availability of congenic strains of mice bearing T cell markers that differ by epitopes that can be identified by monoclonal antibodies allows us to track adoptively transferred cells in a semisyngeneic host. In order to perform quantitative tracking studies of the infused cells, the congenic strain of B6 mouse that expresses CD45.1 can be used to generate TDLN for transfer into CD45.2 hosts. Analysis of the infiltrate can be performed by mechanical dissociation of the tumors in order to recover viable lymphoid infiltrates. By FACS analysis, the number of transferred CD4/CD8 T cells can be quantified. Proliferation of infused cells can be assessed by labeling them with CFSE. Confirmed correlation between effective tumor regression and the infiltration of infused cells to tumor should encourage further attempts to modulate T cell trafficking by biochemical controls or by genetic modification of well-identified adhesion molecules, for example, LFA, ICAM, and selectins. Furthermore, a very recent study described the regulation of T cell trafficking by sphingosine 1-phosphate (S1P) receptor 1 (S1P1) (35). Mature T cells from S1P1 transgenic mice exhibited enhanced chemotactic response toward S1P, and preferentially distributed to the blood rather than secondary lymphoid organs, such as draining lymph nodes. This work suggests that S1P1 affects systemic trafficking of peripheral T cells, and therefore makes the S1P/S1P1 signaling pathway a novel target for T cell trafficking modulation.

IDENTIFICATION AND CHARACTERIZATION OF T CELL SUBSETS RESPONSIBLE FOR ANTITUMOR REACTIVITY

The $CD8^+$ CTLs have long been recognized as the effector cells that mediate tumor regression. In addition, $CD4^+$ effector T cells and NK cells have also been identified to directly or indirectly mediate tumor regression. We reported that CD28 costimulation of tumor-primed lymphoid cells promotes the generation of potent tumor-reactive effector cells, particularly $CD4^+$ T cells. These anti-CD3/anti-CD28 activated $CD4^+$ TDLN cells could independently mediate tumor regression in adoptive immunotherapy (13,14,21).

It has to be presumed that any source of antitumor reactive T cells derived from the tumor-bearing host, that is, TDLN, will represent a small percentage of the total population of retrieved cells. Therefore, a theoretically practical approach would be the identification, isolation, activation and expansion of subsets of T cells capable of mediating tumor regression. In this endeavor, Shu and coworkers found that the down-regulation of the homing molecule L-selectin could serve as a surrogate marker for the isolation of specific tumor-sensitized T cells (18). In adoptive immunotherapy of established intracranial MCA 205 tumors, L-selectin^{low} (CD62L^{low}) cells displayed at least 30-fold greater therapeutic efficacy than unfractionated cells. The L-selectin^{high} cells did not demonstrate any antitumor effects. These results demonstrate that the purification of L-selectin^{low} cells led to the generation of immune effector cells with unusually high therapeutic efficacy against chemically induced tumors. After that, Plautz et al. used advanced tumor models in a stringent comparison of efficacy for the L-selectin^{low} subset versus the total population of TDLN cells following culture in high dose IL-2. L-selectin^{low} subset comprised 5–7% of the TDLN cells. Adoptive transfer of activated L-selectin^{low} cells eliminated 14-day pulmonary metastases and cured 10-day subcutaneous tumors, whereas transfer of maximally tolerated numbers of unfractionated TDLN cells was not therapeutic (36). At the same time, it was identified that tumor-induced L-selectin^{high} cells were suppressor T cells that mediated potent effector T cell blockade and caused failure of otherwise curative adoptive immunotherapy (37). The treatment failure using unfractionated TDLN cells was due to cotransfer of the L-selectin^{high} suppressor T cells present in TDLN. However, the L-selectin^{high} suppressor T cells were only found in day-12 TDLN. In contrast, day-9 TDLN and normal spleens lacked L-selectin^{high} cells.

It was not long before a second surrogate marker was identified for the isolation of tumor-specific T cells. Stoolman et al. described that tumor-specific responses in TDLN were concentrated in cells expressing P-selectin ligand (Plig^{high} T cells) (38). This study found that the minor subset of TDLN T cells expressing binding sites for the adhesion receptor P-selectin (Plig^{high} T cells) produced T lymphoblasts with the most tumor-specific IFNγ synthesis *in vitro* and antitumor activity following adoptive transfer *in vivo*. The cultured Plig^{high} TDLN cells were 10- to 20-fold more active against established pulmonary micrometastases than cultured, unfractionated TDLN, and >30-fold more active than cultured TDLN cells depleted of the Plig^{high} fraction. The Plig^{high} T cells expressed high levels of CD69 and low levels of CD62L (L-selectin^{low}), which agrees with the previous studies on L-selectin in TDLN. Further supporting these observations is a recent study indicating that recruitment of IFNγ-producing cells into the inflamed retina *in vivo* is preferentially regulated by P-selectin glycoprotein ligand (39).

In a different attempt to selectively activate tumor-sensitized T cells, superantigens were utilized *in vitro* to stimulate effector cell generation in a murine model (40). The TDLN cells stimulated with staphylococcal enterotoxins A (SEA), B (SEB) or C2 (SEC2) resulted in the selective expansion of Vβ3 and 11, Vβ3 and 8, or Vβ8.2 T cells, respectively. Adoptive transfer studies revealed that SEB- and SEC2-, but not SEA- stimulated cells mediated tumor-specific regression. These results suggested that T cells bearing Vβ8 may preferentially respond to the growing tumor than T cells bearing Vβ3 or 11 elements of the T cell receptor. Similarly, stimulating TDLN cells with different anti-Vβ mAbs instead of the pan-T cell reagent anti-CD3 mAb enabled the selective activation of Vβ T cell subsets (17). Enrichment of Vβ subsets of TDLN cells revealed that Vβ8^{+} cells released high amounts of IFNγ and GM–CSF with minimal amount of IL-10 in response to tumor, and mediated tumor regression *in vivo*. In contrast, enriched population of Vβ5^{+}, Vβ7^{+}, and Vβ11^{+} cells released low amounts of IFNγ and GM–CSF with high levels of IL-10, and had no *in vivo* antitumor reactivity. *In vitro* depletion of specific Vβ subsets from the whole TDLN pool confirmed that the profile of cytokine released correlated with *in vivo* antitumor function. These studies indicate that functional Vβ subpopulations of effector cells express differential antitumor reactivity, and that selective stimulation of tumor-sensitized T cells is feasible and may represent a more efficient method of generating therapeutic T cells for therapy.

Application of cell subsets for successful T cell therapy should include two approaches: identification of T cell subsets responsible for mediating antitumor reactivity as discussed above, and simultaneously, the elimination of those subsets that are non-reactive or even suppressive. Characterization of regulatory CD4^{+}CD25^{+} T cell subpopulation in terms of their potential suppressive effects on anticancer effector cells would warrant further investigations in this area. A current study showed that CD8^{+} T cell immunity against a tumor self-antigen is augmented by CD4^{+} T helper cells, but hindered by naturally occurring CD4^{+}CD25^{+} T regulatory cells (Treg cells)(41). Adoptive transfer of tumor-reactive CD8^{+} T cells plus CD4^{+}CD25^{-} Th cells into CD4-deficient hosts induced autoimmunity and regression of established melanoma. However, transfer of CD4^{+} T cells that contained a mixture of CD4^{+}CD25^{-} and CD4^{+}CD25^{+} Treg cells or Treg cells alone prevented effective adoptive immunotherapy. These findings thus suggest that adoptive immunotherapy requires the absence of naturally occurring CD4^{+}CD25^{+} Treg cells to be effective, and the optimal composition of a cellular agent should be composed of CD8^{+} cells plus CD4^{+}CD25^{-} cells.

ADOPTIVE T CELL IMMUNOTHERAPY OF CANCER IN LYMPHOPENIC HOST

Studies in the late 1970s demonstrated that the induction of lymphopenia by sublethal total body irradiation can be beneficial for the treatment of tumors in mice (42). Chang et al. reported that the adoptive transfer of immune cells in the irradiated host confers improved therapeutic effects compared to the normal host (43). The role of lymphodepletion on the efficacy of T cell therapy is incompletely understood and may depend on the destruction of CD4^{+}CD25^{+} regulatory cells, interruption of homeostatic T cell regulation, or abrogation of other normal tolerogenic mechanisms. A report by Dummer et al. indicated that the reconstitution of the lymphopenic, sublethally irradiated murine host with syngeneic T cells triggered an antitumor autoimmune response that required expansion within lymph nodes (44).

There are several different animal models of lymphopenia that can be utilized. These include the use of whole body irradiation (WBI), chemotherapy-induced, or genetically altered hosts (i.e., RAG1 knockout mice) that are deficient of T and B cells. The use of various chemotherapeutic agents to induce lymphopenia would simulate the clinical setting. Cyclophosphamide (CTX) is an agent that has been extensively used in murine models and is actively used in the therapy of certain human cancers. It has been described to eliminate tumor-induced suppressor cells in both animal and human settings.

A few years ago, a report described a phase I study of the adoptive transfer of cloned melanoma antigen-specific T lymphocytes for therapy of patients with advanced melanoma (45). Clones were derived from peripheral blood lymphocytes or TILs of patients. Twelve patients received two cycles of cells. Peripheral blood samples were analyzed for persistence of transferred cells by TCR-specific PCR. Transferred cells reached a maximum level at 1 h after transfer, but rapidly declined to undetectable levels by 2 weeks. The lack of clinical effectiveness of this protocol suggested that transfer of different or additional cell types, or that modulation of the recipient host environment was required for successful therapy. Relevant to these studies is the clinical experience reported by Rosenberg and coworkers who infused tumor-reactive T cells in melanoma patients after a nonmyeloablative conditioning regimen (cyclophosphamide/fludarabine) (46). Conditioning with the nonmyeloablative chemotherapy before adoptive transfer of activated tumor-reactive T cells enhanced tumor regression and increased the overall rates of objective clinical responses. Six of thirteen patients demonstrated significant clinical responses as well as autoimmune melanocyte destruction. In a follow up of this experience in 25 patients, the conditioning regimen was given prior to adoptive T cell therapy as before (47). Examination of the T cell persistence through analysis of the specific TCR demonstrated that there was a significant correlation between tumor regression and the degree of persistence in peripheral blood of adoptively transferred T cell clones. Transferred cells persisted for as long as 2 months in the lymphopenic setting induced by the conditioning regimen. In contrast, they presented in the blood for only 2 or 3 weeks without the prior chemotherapy. These series of studies strongly suggest that the lymphopenic host induced by the nonmyeloablative conditioning regimen may provide a better environment for the functioning of the transferred T cells, and hence improve their therapeutic efficacy. Examination of the mechanisms involved in the reconstitution of the lymphodepleted host after adoptive T cell transfer will be important in identifying methods to improve the efficacy of T cell therapies for cancer.

REDIRECT EFFECTOR T CELLS TO TUMOR

As mentioned earlier, the low precursor frequency of tumor-specific T cells in patients hampers routine isolation of these cells for adoptive transfer. To overcome this problem, "targeted adoptive immunotherapy" or "genetic adoptive immunotherapy" has become an attractive option for cancer treatment. This strategy can be approached in two ways: introduction of a chimeric TCR into effector cells; or introduction of a tumor-specific TCR into naïve cells.

The T-body approach uses patient-derived lymphocytes transfected with chimeric receptor genes constructed with the variable domains of monoclonal antibodies or cytokines linked to the constant regions of TCR. The rationale for this novel approach to redirect effector cells combines the effector functions of T lymphocytes with the ability of antibodies or cytokines to recognize predefined surface antigens or cytokine receptors with high specificity and in a non-MHC restricted manner.

Eshhar et al. (48) was one of the first to describe this approach by developing a chimeric receptor gene which recognized trinitrophenyl (TNP). Retroviral transduction of the anti-TNP/TCR chimeric gene into a T cell hybridoma line resulted in gene expression. These gene modified T cells were cytolytic and released IL-2 in response to TNP-labeled Daudi cells, but not unlabeled cells. Also among the pioneers in this area, Hwu et al. (49) developed a recombinant chimeric receptor against an epitope expressed on the majority of ovarian cancer cell lines. The TIL were transduced with this chimeric gene and evaluated for immunologic function. The gene modified TIL showed specific lysis of an ovarian carcinoma cell line, but not nonovarian cell lines. In a direct comparison, the gene modified TIL showed greater therapeutic efficacy *in vivo* than the nontransduced TIL (49). Pinthus et al. evaluated the therapeutic efficacy of anti-erbB2 chimeric receptor-bearing human lymphocytes on human prostate cancer xenografts in a SCID mouse model (50). Local delivery of erbB2-specific transgenic T cells to well-established subcutaneous and orthotopic tumors resulted in retardation of tumor growth and prolongation of animal survival. In a setting of metastatic cancer (51), anti-erbB2 chimeric receptor-modified T cells killed breast cancer cells and secreted IFNγ in an Ag-specific manner *in vitro*. Treatment of established metastatic disease in lung and liver with these genetically engineered T cells resulted in dramatic increases in survival of the xenografted mice. In another report, CD4$^+$ cells isolated from the peripheral blood and engrafted with a recombinant immunoreceptor specific for carcinoembryonic Ag (CEA) efficiently lysed target cells in a MHC-independent fashion, and the efficiency was similar to that of grafted CD8$^+$ T cells (52). In an attempt to further improve the therapeutic utility of redirected T cells, T lymphocytes were transferred with CEA-reactive chimeric receptors that incorporate both CD28 and TCR-zeta signaling domains. T cells expressing the single-chain variable fragment of Ig (scFv)-CD28-zeta chimera demonstrated a far greater capacity to control the growth of CEA$^+$ xenogeneic and syngeneic colon carcinomas *in vivo* compared with scFv-CD28 or scFv-zeta transfected T cells. This study has illustrated the ability of a chimeric scFv receptor capable of harnessing the signaling machinery of both TCR-zeta and CD28 to augment T cell immunity against tumors (53).

In addition to antibodies, cytokines could also be used to reconstruct chimeric TCRs. The IL-13 receptor alpha2 (IL-13Rα2) is a glioma-restricted cell-surface epitope not otherwise detected within the central nervous system. Kahlon et al. (54) described a novel approach for targeting glioblastoma multiforme (GBM) with IL-13Rα2-specific CTLs. The chimeric TCR incorporates IL-13 for selective binding to IL-13Rα2. This represents a new class of chimeric immunoreceptors that signal through an engineered immune synapse composed of membrane-tethered cytokine (IL-13) bound to cell-surface cytokine receptors (IL-13Rα2) on tumors. Human IL-13-redirected CD8$^+$ CTL transfectants display IL-13Rα2-specific antitumor effector function including tumor cell cytolysis and cytokine production. *In vivo*, the adoptive transfer of genetically modified CTL clones resulted in the regression of established human glioblastoma orthotopic xenografts.

The second genetic approach to redirect T cells involves the introduction of tumor-specific TCRs into naïve cells. Genes encoding tumor antigen-specific TCRs can be introduced into primary human T cells as a potential method of providing patients with a source of autologous tumor-reactive T cells. Several tumor-associated antigens have been identified and cloned from human tumors, such as melanoma, breast cancers, and RCC. The antigens have been identified by their ability to induce T cell reactivity by their binding to the TCR $\alpha\beta$ complex. The subsequent cloning of functional TCR genes capable of recognizing tumor-associated antigens offers a potential opportunity to genetically modify naive cells that have not been previously exposed to tumor antigen and to become competent in recognizing tumor. Cole et al. (55) transfected the cDNA for the TCR α and β chains of an HLA-A2 restricted, melanoma-reactive T cell clone into the human Jurkat T cell line. The transfected line was able to mediate recognition of the melanoma antigen, MART-1, when presented by antigen-presenting cells. This represented the first report of a naive cellular construct designed to mediate functional tumor antigen recognition. A recent study explored the simultaneous generation of CD8$^+$ and CD4$^+$ melanoma-reactive T cells by retroviral-mediated transfer of a TCR specific for HLA-A2-restricted epitope of the melanoma antigen tyrosinase

(56). The TCR-transduced normal human peripheral blood lymphocytes secreted various cytokines when stimulated with tyrosinase peptide-loaded antigen-presenting cells or melanoma cells in an HLA-A2-restricted manner. Rosenberg and co-worker (57) isolated the α and β chains of the TCR from a highly avid anti-gp100 CTL clone and constructed retroviral vectors to mediate gene transfer into primary human lymphocytes. The biological activity of transduced cells was confirmed by cytokine production following coculture with stimulator cells pulsed with gp100 peptides, but not with unrelated peptides. The ability of the TCR gene to transfer Ag recognition to engineered lymphocytes was confirmed by HLA class I-restricted recognition and lysis of melanoma tumor cell lines. In addition, nonmelanoma-reactive TIL cultures developed antimelanoma activity following anti-gp100 TCR gene transfer. Together, these studies suggest that lymphocytes genetically engineered to express melanoma antigen-specific TCRs may be of value in the adoptive immunotherapy of patients with melanoma.

The HPV16 (human papilloma virus type 16) infection of the genital tract is associated with the development of cervical cancer in women. The HPV16-derived oncoprotein E7 is expressed constitutively in these lesions and represents an attractive candidate for T cell mediated adoptive immunotherapy. In a recent study, Scholten et al. reported that HPV16E7 TCR gene transfer is feasible as an alternative strategy to generate human HPV16E7-specific T cells for the treatment of patients suffering from cervical cancer and other HPV16-induced malignancies (58). These TCR genes specific for HPV16E7 were isolated and transferred into peripheral blood-derived $CD8^+$ T cells. Biological activity of the transgenic CTL clones was confirmed by lytic activity and IFNγ secretion upon antigen-specific stimulation. Most importantly, the endogenously processed and HLA-A2 presented HPV16E7 CTL epitope was recognized by the TCR-transgenic T cells. In a separate study, ovalbumin (OVA)-specific $CD4^+$ cells were successfully generated. Chamoto et al. (59) prepared mouse antigen-specific Th1 cells from nonspecifically activated T cells after retroviral transfer of TCR genes. These Th1 cells transduced with the α and β genes of the I-A (d)-restricted OVA-specific TCR produced IFNγ in response to stimulation with OVA peptides or A20 B lymphoma cells expressing OVA as a model tumor antigen. The TCR-transduced Th1 cells also exhibited cytotoxicity against tumor cells in an antigen-specific manner. In addition, adoptive transfer of TCR-transduced Th1 cells exhibited potent antitumor activity *in vivo*.

Genetic alteration of T cells with chimeric receptor genes or antigen-specific TCR genes confers the redirection of effector cells to the tumor for its destruction. These approaches may offer novel opportunities to develop immunocompetent effector cellular reagents and improve the efficacy of adoptive immunotherapy of cancer.

COMBINED THERAPY

Cancer is a disease that involves multiple gene malfunctions and numerous biochemical and cellular event errors during its development and metastasis within an individual. Therefore, it is difficult to achieve success utilizing adoptive T cell transfer as a monotherapy. The above-reviewed use of vaccination to induce tumor-reactive pre-effector *in vivo*; the coadministration of immune adjuvant with T cell transfer; and the gene therapy to redirect T cells to tumor are all among the strategies taken to elicit and/or strengthen the efficacy of T cell therapy. Combination therapy is a very common practice during the treatment of diseases. Active vaccine therapy, for example, can be used in concert with chemotherapy, radiotherapy, or antibody therapy. Combining a glioma tumor vaccine engineered to express the membrane form of macrophage colony-stimulating factor with a systemic antiangiogenic drug-based therapy cured rats bearing 7 day old intracranial gliomas (60). We successfully demonstrated that local radiotherapy potentiates the therapeutic efficacy of intratumoral dendritic cell (DC) administration (61), and that anti-CD137 monoclonal antibody administration augments the antitumor efficacy of DC-based vaccines (62).

In order to enhance the efficiency of T cell therapy, various strategies have been employed accompanying cell transfer. These combined therapies include cell transfer in combination with intratumoral expression of lymphotactin (63), DC vaccination (64), or blockade of certain molecules expressed in tumor cells, such as B7-H1 (65).

One of the major obstacles to successful adoptive T cell therapy is the lack of efficient T cell infiltration of tumor. Combined intratumoral lymphotactin (Lptn) gene transfer into SP2/0 myeloma tumors and adoptive immunotherapy with tumor specific T cells eradicated well-established SP2/0 tumors in six of eight mice, and dramatically slowed down tumor growth in the other two mice (63). Cell tracking using labeled T cells revealed that T cells infiltrated better into the Lptn-expressing tumors than non-Lptn-expressing ones. These data provide solid evidence of a potent synergy between adoptive T cell therapy and Lptn gene therapy as a result of facilitated T cell targeting. Dendritic cells are well-known potent antigen-presenting cells. Hwu and co-workers (64) reported that DC vaccination could improve the efficacy of adoptively transferred T cells to induce an enhanced antitumor immune response. Mice bearing B16 melanoma tumors expressing the gp100 tumor antigen were treated with activated T cells transgenic for a TCR specifically recognizing gp100, with or without concurrent peptide-pulsed DC vaccination. Antigen-specific DC vaccination induced cytokine production, enhanced cell proliferation, and increased tumor infiltration of adoptively transferred T cells. The combination of DC vaccination and adoptive T cell transfer led to a more robust antitumor response than the use of each treatment individually. This work shows that in addition to their ability to initiate cell-mediated immune responses by stimulating naive T cells, dendritic cells can strongly boost the antitumor activity of activated T cells *in vivo* during adoptive immunotherapy. Certain cell surface molecules, expressed either on tumor cells or on T cells, have demonstrated have demonstrated potential suppressive impact on the adoptive T cell immunotherapy. For example, during the last few years, new members of the B7 family molecules have been identified, for example, B7-H1, which is constitutively expressed on 66% of freshly isolated squamous cell carcinomas of the head and neck (SCCHN) (65). When B7-H1-negative mouse SCC line, SCCVII, was transfected to express B7-H1, all of the animals succumbed to B7-H1/SCCVII tumors even after adoptive T cell immunotherapy. However, the infusion of B7-H1 blocking monoclonal antibody with activated T cells cured 60% of animals. The data support B7-H1 blockade as a new approach to enhance the efficacy of T cell immunotherapy. These findings also illuminate a new potential application for the blockade of certain "negative costimulation molecules" on T cells, for example, CTLA-4 and programmed death-1 (PD-1) molecules. This kind of blocking may augment the therapeutic efficacy mediated by the transferred T cells. The blockade can be done using specific monoclonal antibodies, soluble ligands for CTLA-4 or PD-1, or by synthesized antagonists. In addition, effector cells can be derived from the animals deficient in the relevant molecules for preclinical investigations.

Immune tolerance of tumor-bearing host represents another major obstacle for the successful use of adoptive T cell immunotherapy. A recent study examined the requirement for assistance to the low affinity tumor-specific $CD8^+$ T cells transferred into tumor-bearing mice (66). The TCR transgenic mice expressing a class I-restricted hemagglutinin (HA)-specific TCR (clone 1 TCR) were generated. Upon transfer into recipient mice in which HA is expressed at high concentrations as a tumor-associated Ag, the clone 1 TCR $CD8^+$ T cells exhibited very weak effector function and were soon tolerized. However, when HA-specific $CD4^+$ helper cells were co-transferred with clone 1 cells and the recipients were vaccinated with influenza, clone 1 cells were found to exert a significant level of effector function and delayed tumor growth. This work shows that in order to optimize the function of low avidity tumor-specific T cells after adoptive transfer, additional measures need to be taken to help break the host tolerance.

Effective tumor therapy requires a proinflammatory microenvironment that permits T cells to extravasate and to destroy the tumor. Proinflammatory environment can be induced by various chemical, physical, and immunological protocols. Greater extent of success can be expected by combining adoptive T cell therapy with the traditional cancer treatment methods, for example, surgery, chemotherapy, and radiation therapy, as well as with different forms of immunotherapeutic strategies, such as vaccine, antibody, cytokines, gene therapy, and so on. The factors to be combined can involve two or more approaches.

In summary, adoptive immunotherapy utilizing tumor-reactive T cells offers a promising alternative approach for the management of cancer. Through the endeavors of clinical and basic research scientists during the last two decades, the process of adoptive T cell therapy of cancer has evolved from its original single-step approach into its current multiple-step procedure. Successful T cell immunotherapy of cancer is the outcome of this multi-step process that depends on successful Ag priming, numerical amplification of low frequency Ag-specific precursors, use of immune adjuvants, and efficient infiltration of tumors in all metastatic sites by effector T cells. New directions in this field include the identification and application of tumor-reactive subpopulation of T cells, creation of a lymphopenic environment in the recipient host, and the redirection of the effector cells toward the tumor. Development of these latter techniques and the combined use of different therapeutic strategies may further improve the efficacy of the immunotherapy of human cancer employing adoptive T cell transfer. Studies and developments of immunotherapy for cancer should accelerate the application of this strategy in infectious disease, autoimmune disease and other disease managements.

BIBLIOGRAPHY

1. Chang AE, et al. Adoptive Immunotherapy with vaccine-primed lymph node cells secondarily activated with anti-CD3 and interleukin-2. J Clin Oncol 1997;15:796.
2. Chang AE., et al. Generation of vaccine-primed lymphocytes for the treatment of head and neck cancer. Head and Neck 2003;25:198.
3. Chang AE, et al. Phase II trial of autologous tumor vaccination, Anti-CD3-activated vaccine-primed lymphocytes, and Interleukin-2 in stage IV renal cell cancer. J Clin Oncol 2003;21:884.
4. Shu S, Chou T, Rosenberg SA. Generation from tumor-bearing mice of lymphoid cells with in vivo therapeutic efficacy. J Immunol 1987;139:295–304.
5. Chou T, Chang AE, Shu S. Generation of therapeutic T lymphocytes from tumor-bearing mice by in vitro sensitization: Culture requirements and characterization of immunologic specificity. J Immunol 1988;140:2453–2461.
6. Geiger J, Wagner P, Shu S, Chang AE. A novel role for autologous tumor cell vaccination in the immunotherapy of the poorly immunogenic B16-BL6 melanoma. Surg Oncol 1992;1:199–208.
7. Geiger J, et al. Generation of T- cells reactive to the poorly immunogenic B16-BL6 melanoma with efficacy in the treatment of spontaneous metastases. J Immunother 1993;13:153–65.
8. Li Q, et al. Immunological effects of BCG as an adjuvant in autologous tumor vaccines. Clin Immunol 2000;94:64–72.
9. Restifo N, et al. A nonimmunogenic sarcoma induced with the cDNA for interferon-γ elicits $CD8^+$ T cells against the wild-type tumor: Correlation with antigen presentation capability. J Exp Med 1992;175:1423–1431.
10. Chang AE, et al. Immunogenetic therapy of human melanoma utilizing autologous tumor cells transduced to secrete GM–CSF. Hum Gene Ther 2000;11:839–850.
11. Liu J, et al. Ex vivo activation of tumor-draining lymph node T cells reverses defects in signal transduction molecules. Can Immunol Immunother 1998;46:268.
12. Yoshizawa H, Chang AE, Shu S. Specific adoptive immunotherapy mediated by tumor-draining lymph node cells sequentially activated with anti-CD3 and IL-2. J Immunol 1991;147:729–37.
13. Li Q, Furman SA, Bradford CR, Chang AE. Expanded tumor-reactive $CD4^+$ T-Cell responses to human cancers induced by secondary anti-CD3/anti- CD28 activation. Clin. Can. Res. 1999;5:461.
14. Li Q, et al. Therapeutic effects of tumor-reactive $CD4^+$ cells generated from tumor-primed lymph nodes using anti-CD3/anti-CD28 monoclonal antibodies. J Immunother 2002;25:304.
15. Ito F, et al. Antitumor therapy reactivity of Anti-CD3/Anti-CD28 bead-activated lymphoid cells: Implications for cell in a murine model. J Immunother 2003;26:222.

16. Li Q, et al. Polarization effects of 4-1BB during CD28 costimulation in generating tumor-reactive T Cells for cancer immunotherapy. Can. Res. 2003;63:2546.
17. Aruga A, et al. Type 1 versus type 2 cytokine release by Vβ T cell subpopulations determines in vivo antitumor reactivity: IL-10 mediates a suppressive role. J Immunol 1997;159:664–673.
18. Kagamu H, Shu S. Purification of L-selectin^{low} cells promotes the generation of highly potent CD4 antitumor effector T lymphocytes. J Immunol; 1998;160:3444–3452.
19. Hart-Meyers J, et al. Cutting Edge: CD94/NKG2 is expressed on Th1 but not Th2 cells and costimulates Th1 effector functions. J Immunol 2002;169:5382–5386.
20. Hou W, et al. Pertussis toxin enhances Th1 responses by stimulation of dendritic cells. J Immunol 2003;170:1728–1736.
21. Li Q, et al. Synergistic effects of IL-12 and IL-18 in skewing tumor-reactive T-cell responses towards a type 1 pattern. Can. Res. 2005;65:1063.
22. Topalian SL. et al. Immunotherapy of patients with advanced cancer using tumor-infiltrating lymphocytes and recombinant interleukin-2: a pilot study. J. Clin. Oncol. 1988;6:839.
23. Shrikant P, Mescher MF. Opposing effects of IL-2 in tumor immunotherapy: promoting CD8 T cell growth and inducing apoptosis. J. Immunol. 2002;169:1753.
24. Rosenberg SA, et al. Experience with the use of high-dose IL-2 in the treatment of 652 cancer patients. Ann Surg 1989;210:474.
25. Roychowdhury S, et al. Failed adoptive immunotherapy with tumor-specific T cells: Reversal with low-dose interleukin 15 but not low-dose interleukin 2. Can Res 2004;64:8062.
26. Nacsa J, et al. Contrasting effects of low-dose IL-2 on vaccine-boosted Simian Immunodeficiency Virus (SIV)-specific CD4$^+$ and CD8$^+$ T cells in macaques chronically infected in SIVmac251. J Immunol 2005;174:1913.
27. Poggi A, et al. Tumor-induced apoptosis of human IL-2-activated NK cells: role of natural cytotoxicity receptors. J Immunol 2005;174:2653.
28. Yajima T, et al. A novel role of IL-15 in early activation of memory CD8$^+$ CTL after reinfection. J Immunol 2005;174:3590–3597.
29. Kamiya S, et al. An indispensable role for STAT1 in IL-27-induced T-bet expression but not proliferation of naïve CD4$^+$ T cells. J Immunol 2004;173:3871–3877.
30. Osaki T, et al. IFN-γ-inducing factor/IL-18 administration mediates IFN-γ-and IL-12-independent antitumor effects. J Immunol 1998;160:1742.
31. Palmer DC, et al. Vaccine-stimulated adoptively transferred CD8$^+$ T cells traffic indiscriminately and ubiquitously while mediating specific tumor destruction. J Immunol 2004;173:7209–7216.
32. Mukai S, Kjaergaard J, Shu S, Plautz GE. Infiltration of tumors by systemically transferred tumor-reactive T lymphocytes is required for antitumor efficacy. Can Res 59: 5245.
33. Skitzki J, et al. Donor cell cycling, trafficking, and accumulation during adoptive immunotherapy for murine lung metastases. Can Res 2004;64:2183.
34. Meidenbauer N, et al. Survival and tumor localization of adoptively transferred melan-A-specific T cells in melanoma patients. J Immunol 2003;170:2161–2169.
35. Chi H, Flavell RA. Cutting Edge: Regulation of T cell trafficking and primary immune responses by sphingosine 1-phosphate receptor 1. J Immunol 2005;174:2485–2488.
36. Wang LX, Chen BG, Plautz GE. Adoptive immunotherapy of advanced tumors with CD62 L-selectin^{low} tumor-sensitized T lymphocytes following ex vivo hyperexpansion. J Immunol 2002;169:3314–3320.
37. Peng L, et al. Tumor-induced L-selectinhigh suppressor T cells mediate potent effector T cell blockade and cause failure of otherwise curative adoptive immunotherapy. J Immunol 2002;169:4811–4821.
38. Tanigawa K, et al. Tumor-specific responses in lymph nodes draining murine sarcomas are concentrated in cells expressing P-selectin binding sites. J Immunol 2001;167:3089–3098.
39. Xu H, et al. Recruitment of IFN-γ-producing (Th1-like) cells into the inflamed retina in vivo is preferentially regulated by P-selectin glycoprotein ligand 1:P/E-selectin interactions. J Immunol 2004;172:3215–3224.
40. Shu S, et al. Stimulation of tumor-draining lymph node cells with superantigenic staphyloccocal toxins leads to the generation of tumor-specific effector cells. J Immunol 1994;152:1277–1288.
41. Antony PA, et al. CD8$^+$ T cell immunity against a tumor/self-antigen is augmented by CD4$^+$ T helper cells and hindered by naturally occurring T regulatory cells. J Immunol 2005;174:2591–2601.
42. Mule JJ, Jones FR, Hellstrom I, Hellstrom KE. Selective localization of radio-labeled immune lymphocytes into syngeneic tumors. J Immunol 1979;123:600.
43. Chang AE, et al. Differences in the effects of host suppression on the adoptive immunotherapy of subcutaneous and visceral tumors. Can Res 1986;46:3426.
44. Dummer W, et al. T cell homeostatic proliferation elicits effective antitumor autoimmunity. J Clin Inves 2002;110:185.
45. Dudley ME, et al. Adoptive transfer of cloned melanoma-reactive T lymphocytes for the treatment of patients with metastatic melanoma. J Immunoth 2001;24:363–373.
46. Dudley ME, et al. Cancer regression and autoimmunity in patients after clonal repopulation with antitumor lymphocytes. Science 2002;298:850.
47. Robbins PF, et al. Cutting Edge: Persistence of transferred lymphocyte clonotypes correlates with cancer regression in patients receiving cell transfer therapy. J Immunol 2004;173:7125–7130.
48. Eshhar Z, Waks T, Schindler DG, Gross G. Specific activation and targeting of cytotoxic lymphocytes through chimeric single chains consisting of antibody binding domains and the g or z subunits of the immunoglobulin and T cell receptors. Proc Natl Acad Sci USA 1993;90:720–724.
49. Hwu P, et al. In vivo antitumor activity of T cells redirected with chimeric antibody/T cell receptor genes. Can Res 1995;55:3369–3373.
50. Pinthus JH, et al. Immuno-gene therapy of established prostate tumors using chimeric receptor-redirected human lymphocytes. Can Res 2003;63:2470–2476.
51. Kershaw MH, et al. Gene-engineered T cells as a superior adjuvant therapy for metastatic cancer. J Immunol 2004;173:2143–2150.
52. Hombach A, et al. CD4$^+$ T cells engrafted with a recombinant immunoreceptor efficiently lyse target cells in a MHC antigen-and-fas independent fashion. J Immunol 2001;167:1090–1096.
53. Haynes NM, et al. Rejection of syngeneic colon carcinoma by CTLs expressing single-chain antibody receptors codelivering CD28 costimulation. J Immunol 2002;169:5780–5786.
54. Kahlon KS, et al. Specific recognition and killing of glioblastoma multiforme by interleukin 13-zetakine redirected cytolytic T cells. Can Res 2004;64:9160–9166.
55. Cole DJ, et al. Characterization of the functional specificity of a cloned T cell receptor heterodimer recognizing the MART-1 melanoma antigen. Can Res 1995;55:748–752.
56. Roszkowski JJ, et al. Simultaneous generation of CD8$^+$ and CD4$^+$ melanoma-Reactive T cells by retroviral-mediated transfer of a single T-cell receptor. Can Res 2005;65:1570–1576.

57. Morgan RA, et al. High efficiency TCR gene transfer into primary human lymphocytes affords avid recognition of melanoma tumor antigen glycoprotein 100 and does not alter the recognition of autologous melanoma antigens. J Immunol 2003;171:3287–3295.
58. Scholten KB, et al. Preservation and redirection of HPV16E7-specific T cell receptors for immunotherapy of cervical cancer. Clin Immunol 2005;114:119–129.
59. Chamoto K, et al. Potentiation of tumor eradication by adoptive immunotherapy with T-cellreceptor gene-transduced T-helper type 1 cells. Can Res 2004;64:386–390.
60. Jeffes EWB, et al. Antiangiogenic drugs synergize with a membrane macrophage colony-stimulating factory-based tumor vaccine to therapeutically treat rats with an established malignant intracranial glioma. J Immunol 2005;174:2533–2543.
61. Teitz-Tennenbaum S. et al. Radiotherapy potentiates the therapeutic efficacy of intratumoral dendritic cell administration. Can Res 2003;63:8466–8475.
62. Ito F, et al. Anti-CD137 monoclonal antibody administration augments the antitumor efficacy of dendritic cell-based vaccines. Can Res 2004;64:8411.
63. Huang H, Li F, Gordon JR, Xiang J. Synergistic enhancement of antitumor immunity with adoptively transferred tumor-specific CD4$^+$ and CD8$^+$ T cells and intratumoral lymphotactin transgene expression. Can Res 2002;62:2043.
64. Lou Y, et al. Dendritic cells strongly boost the antitumor activity of adoptively transferred T cells in vivo. Can Res 2004;64:6783.
65. Strome SE. B7-H1 blockade augments adoptive T-cell immunotherapy for squamous cell carcinoma. Can Res 2003;63: 6501.
66. Lyman MA, et al. The fate of low affinity tumor-specific CD8$^+$ T cells in tumor-bearing mice. J Immunol 2005;174:2563–2572.

See also BORON NEUTRON CAPTURE THERAPY; MONOCLONAL ANTIBODIES.

IMPEDANCE PLETHYSMOGRAPHY

HELMUT HUTTEN
University of Technology
Graz, Australia

INTRODUCTION

Plethysmography is a volumetric method, that is, a method for the assessment of a volume (the Greek words *plethys* and *plethora* mean full and fullness, respectively). Impedance plethysmography is based on the measurement of passive electrical properties of biological tissues. Those passive electrical properties are parameters of the so-called bioimpedance. The first publication about impedance plethysmography by Nyboer et al. (1) dates back to 1943. Pioneering contributions to the basic understanding of the relations between the assessment of volumes by impedance plethymosgraphy and the electrical properties of biological tissue have been provided by Schwan et al. (2) already in 1955. But already by the end of the nineteenth century Stewart had used the recording of electrical conductivity to study transit times between different sites of the body after injection of saline into the circulation (3). Blood flow recording is one of the most relevant fields for the clinical application of impedance plethysmography nowadays.

Impedance plethysmography is a volumetric method that aims to assess a volume or changes of a volume. Usually, a volume is the filling volume of a space that is enclosed by geometric boundaries. In this case, volumetry means the determination of the boundaries with subsequent assessment of the volume within the boundaries. Those boundaries can be determined by the impedance method if the electrical properties of the substances on both sides of the boundaries are different.

Impedance plethysmography, however, can also be applied to the assessment of volumes that are not lumped compartments within geometric boundaries, for example, it can be used for the volumetric measurement of a certain component within a mixture. Such components may be cells (e.g., the volume of cells in blood), tissues (e.g., the volume of fat tissue in the body), spaces with different composition (e.g., intra- and extracellular spaces), or the volume of the air that is enclosed in the alveoli of lung tissue. In that case, volumetry means the estimation of the space that would be occupied by the respective component if it would be concentrated in one single lumped compartment. Usually, this volume is estimated as a percentage of the whole distribution volume, for example, the volume of cells in blood or the content of water in the whole body. The electrical properties of the respective component must be different from those of all other components. The volumetric assessment does not require a homogeneous distribution of the considered component within the given space if the actual distribution can be taken into account, for example, by a model. Under certain conditions, a tissue can be identified by specific features like morphological structure and/or chemical composition if those features are related with its electrical properties.

The typical application of plethysmography in clinical routine is the diagnosis of those diseases for which the measurement of volumes or changes of volume renders possible the interpretation of functional disorders or functional parameters. The most widely and routinely applied diagnostic examinations are concerned with:

1. Heart: Cardiac mechanical disorders by impedance cardiography (i.e., pumping insufficiency by measuring cardiac stroke volume, including heart rate and other cardiac parameters like ejection period). This application is discussed in another article.
2. Peripheral circulation: Vascular disorders by impedance rheography (i.e., deep venous thrombosis by impedance phlebography, and estimation of blood flow in the brain or other peripheral vessels).
3. Lung: Ventilatory disorders by impedance pneumography (i.e., insufficient ventilation by monitoring the tidal volume and/or respiratory rate).

METHODOLOGY

Impedance plethysmography is a noninvasive method that employs contacting, usually disposable electrodes, in most cases metal-gel electrodes, for example, with Ag/AgCl for the metal plate. Usually, the electrodes have circular geometry; however, other geometries might be preferable, for example, band-like geometry for segmental measurement at the extremities. It must be considered that the metal plates of electrodes are areas with the same potential, and therefore may affect the electromagnetic field distribution in the considered object. Electrodes with small areas help to reduce that effect, whereas electrodes with large areas render it possible to reach a more homogenous current field in the measured object.

Contacting electrodes can easily be attached to the skin or surface of the measurement object, usually by an adhesive material that is already fixed to the electrode. Only in special cases, for example, for research purposes, does the measurement require invasive application.

Different contactless measurement modes gain increasing attention, for example,

1. Microwave-based methods with antennas as applicators and measurement of the scattered electromagnetic field.
2. Methods based on exploiting the magnetic instead of the electrical properties: inductive plethysmography that uses coils and records the changes in the inductance and magnetic susceptibility plethysmography that employs strong magnetic fields and records the changes in the magnetic flux, for example, by superconducting quantum interference devices (SQUID).

All bioimpedance-based methods are aiming at recording either the effect on the applied electromagnetic field by the measurement object or the response of the measurement object to the application of the electromagnetic field. The directly measured quantities are electrical quantities, for example, voltages or currents. Those quantities are actually imaging electrical coefficients, for example, conductivity, permittivity, or resistivity, which are material-specific parameters of the tissue impedance and monitor its morphological structure and/or chemical composition. With these material-specific parameters the geometric boundaries are determined and used for the estimation of the volume or volume changes. Relations between the measured electrical parameters and the volume are usually based on models. The employed models can be very simple, for example, described by simple geometric boundaries like cylinders, spheres, or ellipsoids. More complex 3D models may be described by the finite element method (FEM) or similar approaches, which allows simulating the distribution of the electromagnetic field in the measured object, that is, the current pathways and the iso-potential planes. Sophisticated iterative optimization procedures are employed to match the simulated values with the measured ones (4).

Electrical impedance tomography (EIT) is a direct approach for determining the 3D geometry of those compartments with the same material-specific parameters in a biological object like the human torso or an extremity. This technique uses multiple electrodes, usually arranged in a plane. Mapping (or imaging) of the impedance distribution in the examined cross-sectional layer requires the solution of the inverse or back-projection problem. Actually, the obtained 2D image is a pseudo-3D image since the current pathways are not constrained to the examined layer. Electrical impedance tomography supplies a comparable near-anatomic cross-sectional image comparable to those of other CT-based procedures [e.g., X-ray CT, NMR, or ultrasound (US)], however, with very poor spatial resolution. Boundaries of compartments with the same material-specific coefficients are found by segmentation. Segmented areas with assumed thickness of the single layers are used for volume estimation. Changes in the volume can be assessed by comparing the volumes obtained in consecutive images.

It is a characteristic feature of all methods that record the electrical bioimpedance that the evoked response depends on the strength of the local electromagnetic field. For this reason, it has to be taken into account that the resulting current density may be inhomogeneous in the considered tissue volume. Causes for such nonhomogeneity may be geometric constraints (e.g., a nonregular shape of the considered volume); the composition of the tissue within the considered volume that may be a mixture of tissues with different electrical properties (e.g., blood with low resistivity, or bone with high resistivity, as compared with skeletal muscle). Those different tissues may electrically be switched in parallel or serial order; and the size and the location of the current-feeding electrodes. The current density is higher in regions near to the feeding electrodes than in distant regions. Consequently, the regions near to the feeding electrodes give the strongest contribution to the measured voltage.

This requires (1) careful selection of the current-feeding site. In the tetrapolar mode also the position of the voltage-sensing electrodes must be taken into account; and (2) appropriate consideration of the inhomogeneous distribution of the electrical parameters within the considered tissue volume, that is, the course of blood vessels.

Special electrode arrangements have been developed for certain applications in order to minimize the measurement errors. Concentric multielectrode arrangements with the outer electrodes on a potential different from that of the inner electrode have been proposed with the objective to optimize the current distribution in the measured volume.

The frequency that can be used for the measurement of the passive electrical properties of biological tissue ranges from very low frequencies to some gigahertz. The most popular frequency band for impedance plethysmography is between 1 kHz and 10 MHz. This frequency band encloses the so-called β-dispersion, which is actually a dielectric or structural relaxation process. The β-dispersion is also known as Maxwell–Wagner relaxation. It is characterized by a transition in the magnitude of the electrical parameters with frequency. This transition is caused by the fact that cellular membranes have high impedance below and low impedance above that β-dispersion. For frequencies distinctly below the β-dispersion, the current flow is restricted to the extracellular space. For frequencies distinctly above the β-dispersion, the current can pass through the cellular membrane. Consequently, with frequencies distinctly below the β-dispersion only the volume or volume changes of the extracellular space will be monitored, whereas with frequencies distinctly above the

Table 1. Compilation of Typical Values of Resistivity (Ω·m) of Various Body Tissues[a]

	Frequency							
	10 Hz	100 Hz	1 kHz	10 kHz	100 kHz	1 MHz	10 MHz	100 MHz
Muscle, skeletal	9.6	8.8	8.1	7.6	2.0	1.8	1.6	1.4
Muscle, heart	9.6	9.3	8.0	6.0	2.1	2.0	1.6	1.5
Liver	10.0	8.7	8.6	7.6	4.6	2.8	2.8	1.7
Kidney					1.9	1.8	1.4	1.3
Brain			6.1		6.0	5.3	3.7	1.5
Fatty tissue			23.2					
Blood			1.6	1.5	1.5	1.4	0.9	0.8

[a] Note that those values must not be assumed to represent exact figures since they do not consider important details like species, preparation of sample, time after excision, temperature, or the procedure and protocol for their measurement. The values are compiled from many different sources and, if necessary transformed to resistivity.

β-dispersion, the total volume (i.e., both extra- and intracellular space) or changes of this total volume can be recorded. Using at least one frequency below and another one above the β-dispersion allows determining the ratio of extra- and intracellular spaces, and hence also fluid shifts between these spaces.

Special applications of this approach are the monitoring of fluid exchange processes during hemodialysis (5) and orthostatic challenges (6), the control and management of fluid infusion therapy, the detection of lung edema, and the viability surveillance of organs after blood flow has been stopped during surgery or when the organs are preserved for transplantation (7,8). The viability surveillance is based on the fact that oxygen deficiency with the subsequent lack of energy-rich substrates causes a failure of the active transmembraneous ionic transport mechanisms and, as a consequence, leads to an intracellular edema (i.e., an increase of the intracellular volume). This approach has also been investigated for graft rejection monitoring.

The passive electrical properties are specific for each tissue. They are mainly depending on the content of water, the ratio of extra- and intracellular space, the concentration of electrolytes, and the shape of the cells and their orientation in the electrical field (e.g., of the fibers of skeletal and cardiac muscle). Table 1 shows a compilation of typical values of resistivity of various body tissues. It must be taken into account, however, that these values do not represent exact figures. Exact figures need detailed information about species, preparation of the sample, time after excision, measurement temperature, the employed method, and the protocol for the measurement. Comprehensive data compilations with the supplement of those details are found in Refs. 9 and 10.

These tissue-specific properties can be used for special applications, such as the analysis of the tissue composition or for tissue characterization by Impedance Spectroscopy. Those methods are the subject of another article and will not be discussed here in detail. A very popular application is the determination of total body water (11) or of whole body composition, for example, the determination of the percentage of body fat in order to support adequate nutrition management or control of physical exercises. Such approaches aim for the estimation of the compartmental volume of a certain tissue (e.g., fat) that is mixed with another tissue (e.g. fat-free tissue) in a common space (i.e., the body or an extremity).

FUNDAMENTALS OF BIOIMPEDANCE MEASUREMENT

The most important principle for bioimpedance measurements is the adequate modeling of the passive electrical behavior of the tissue by an equivalent electrical circuit. The validity of simple models is restricted to narrow frequency ranges (e.g., the β-dispersion) and/or to simple geometric shapes of the biological object (e.g., cylinders as a model for extremities). The most widely accepted models for the bioimpedance in the frequency range around the β- dispersion are the *RC*-networks shown in Fig. 1. These models represent the spatially distributed electrical properties by discrete components. Actually, they are only simplified 2D models. The network shown in Fig. 1a is mimicking the biological system and its histological structure. It represents both the extracellular and intracellular space by the resistors R_e and R_i, respectively, and the cell membrane by the capacitor C_m. Since the current passes twice the membrane when flowing through the cell, the two capacitors C_m* in series with R_i can equivalently be expressed by one single capacitor C_m in series with R_i. This network is usually replaced by the one shown in Fig. 1b in which R_s is arranged in series with the parallel circuit of R_p and C_p. These components have no relation with real histological structures. The parameter R_s corresponds to the parallel circuitry of R_e and R_i as can be demonstrated for high frequencies. The parameter R_s can be considered to be very small as compared with R_p. In many cases, R_S may

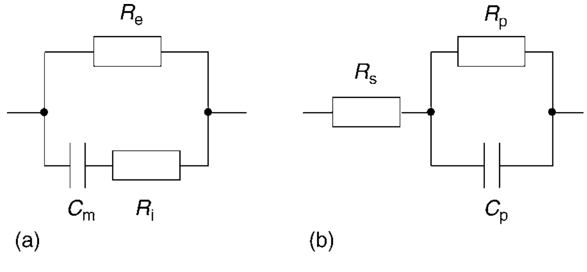


Figure 1. RC-networks modeling tissue impedance. The model in (a) mimicks morphological structures, whereas the model in (b) shows the electrically equivalent, but, more usual circuitry.

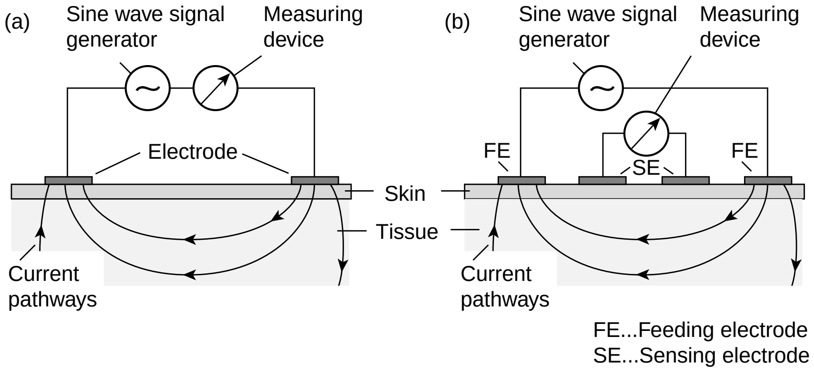


Figure 2. Bioimpedance measurement set-up. (a) Shows the two-electrode configuration, whereas the four-electrode configuration is depicted in (b).

even be neglected for cases of simplification, that is, the electrical model is simply a parallel circuit of a resistor and a capacitor (e.g., $R_p \| C_p$).

When using contacting electrodes, different approaches are possible for the measurement of the bioimpedance. The most important feature is the number of employed electrodes, usually two or four electrodes. More than 4 electrodes, up to 128 electrodes, are primarily used for CT-based impedance imaging like EIT. The two-electrode configuration is called the bipolar mode, and the four-electrode configuration is the tetrapolar mode (Fig. 2). The bipolar mode can be compared with the usual method for impedance measurement by feeding a current into the measurement object and recording the voltage or vice versa. In the tetrapolar mode, two electrodes are the feeding electrodes (usually the outer electrodes) and the other two electrodes are the sensing electrodes (usually the inner ones). In the tetrapolar mode, more than two sensing electrodes can be employed, for example, if monitoring of serial segments at the extremities are to be achieved.

The interface between the electrode with the metallic plate on the one side and the electrolyte on the other side is the boundary where a current carried by electrons is transformed into a current carried by ions. The electrolyte may either be contained in the gel of the electrodes or be the electrolytic fluid in the tissue. The basic process of the charge transfer from electrons to ions is a chemical reaction (12). The simplest model of such an interface is again an impedance consisting of a parallel circuit with a resistor R_F (the Faraday resistance) and a capacitor C_H (the Helmholtz capacitance), i.e., $R_F \| C_H$ (Fig. 3b). Real electrodes show a polarization effect that is caused by a double layer of opposite charges at the interface, actually the Helmholtz capacitance (Fig. 3a). Therefore, the electrode model with $R_F \| C_H$ has to be supplemented with an additional voltage source E_P. The steady-state condition is reached if the tendency of metallic ions to enter the electrolyte and leave behind free electrons is balanced by the electrostatic voltage originating from the double layer. After disturbances, for example, by charge transfer forced by an externally applied voltage, another equilibrium for the double-layer voltage is reached with a time constant depending on R_F and C_H. All these components may have poor stability with time, especially in the period immediately after attaching the electrode on the skin. For surface electrodes, it must also be taken into account that the impedance of the skin, especially the stratum corneum, which is the outmost epidermal layer, can be much larger than the impedance of the deeper tissue (e.g., skeletal muscle), which is in a complex parallel-serial arrangement with the skin. Sweating underneath the electrode lowers the electrode-tissue transimpedance. For the measurement of the impedance of deeper tissues the adequate preparation of the skin by abrasion, stripping, or puncturing at the site of the electrodes might be necessary in order to diminish the transimpedance. This transimpedance depends on the size of the electrode (i.e., the contacting area) and the measurement frequency, and ... additionally on the pressure with which the electrode is attached to the skin. For electrodes, a typical value for the transimpedance is ~100–200 $\Omega \cdot cm^2$.

In the bipolar mode, the two electrode–electrolyte interfaces are in series with the actual bioimpedance of the measured object. Therefore, the recorded impedance is always the sum of at least three impedances. The impedance of the biological sample cannot be calculated as an individual quantity from the recorded impedance value. This is the most serious shortcoming of the bipolar mode.

In the tetrapolar mode, the electrode–electrolyte interface usually can be neglected if the measurement is performed with a device with high input impedance (i.e., with very low current passing across the electrode–tissue inter-

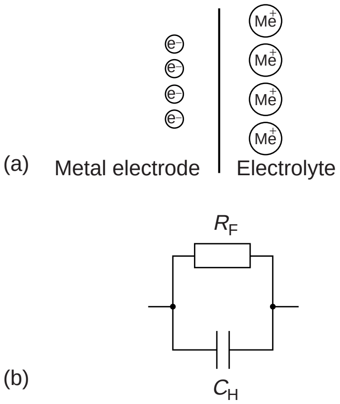


Figure 3. Metal electrode: electrolyte interface. (a) Illustrates the steady-state condition at the boundary. Following the tendency of metallic ions (Me^+) to enter the electrolyte and to leave behind free electrons, the steady state is reached when this tendency is balanced by the electrostatic voltage originating from the double layer. (b) Shows the simplified model with only the passive electrical components, that is, the Faraday resistance R_F and the Helmholtz capacitance C_H.

face). A drawback of the tetrapolar mode, however, is that the measured impedance value cannot be exactly assigned with a certain volume of the tissue between the sensing electrodes, even if the sensing electrodes are positioned on a straight line connecting the two feeding electrodes. Band electrodes that are attached at the whole circumference (e.g., at the extremities), yield more valid results. If circular electrodes are applied and both the feeding and the sensing electrodes are placed on the circumference of a cross-section (e.g., around the thorax or the head), it is nearly impossible to assign the actually measured impedance value with a certain volume within that cross-section due to the complex current field. In those cases, the monitored volume consists of a serial-parallel circuitry of different tissue layers with different electrical properties (e.g., skin, subcutaneous soft tissue, bone, deeper tissues). For this reason, the result of ventilation measurements with electrodes, either disk or band electrodes, placed on the thorax may be affected by the higher conductivity of extra-thoracic muscles as compared with the very low conductivity of the rib cage, which prevents the entrance of current into the pulmonary tissue that is actually the object of interest. Sweating may additionally cause another low impedance parallel circuit along the skin and, thus, yield considerable measurement errors. The situation is similar for the measurement of brain parameters, for example, brain–blood flow or brain edema, with electrodes placed on the scalp. Since the conductivity of the extracranial soft tissue (e.g., skin, muscle) is much higher than the conductivity of the bony skull, only few current pathways will pass through the brain.

The different instrumental approaches for measuring the bioimpedance in the frequency range of the β-dispersion are the impedance bridge (e.g., an extension of the classical Wheatstone bridge); the self-balanced active bridge; the resonance method that is mainly a compensation method; the pulse method that is a time domain procedure and not widely used; the voltage-current method, based either on feeding a constant voltage (i.e., from a generator with low source impedance) and monitoring the resulting current or on feeding a constant current (i.e., from a generator with a sufficiently high source impedance, ~ 100 kΩ since the load impedance may be up to 1 kΩ) and monitoring the resulting voltage. If employing the bipolar mode, it would be more correct in this case to use the term transimpedance than impedance for the actually measured quantity between the electrodes.

For the tetrapolar configuration only the voltage-current method is applicable. Phase angle measurements employ an ohmic resistor in series with the measuring object as reference. Absolute phase angle measurements are questionable even for the bipolar configuration since the measured phase angle always includes the phase shifts that are caused by the two electrode–skin contacts and depend on the actual values of the Faraday resistance and the Helmholtz capacitance. If the Faraday resistance is small and the Helmholtz capacitance is fairly large, the phase shift by the electrode–skin interface may become negligible. This is one of the advantages of electrodes with artificially increased surfaces, for example, electrodes with porous or fractally coated surfaces that might be obtained by sputtering or chemical processes, as compared with polished surfaces.

Usually, the measurement is performed with a constant-voltage generator for technical reasons. The applied feeding signal, whether voltage or current, should be sinusoidal with a very low distortion factor (i.e., with a low content of harmonics) and with high stability both in amplitude and frequency. Any modulation of the feeding signal may provoke an additional response for this undesired modulation frequency that has to be avoided with regard to the frequency dependence of the impedance specific variables. The voltage amplitude is in the range of some volts, the current amplitude is in the range of some microamps to milliamps (μA to mA). The changes in the impedance caused by volume changes can be very small, $<0.001\%$. This means that very small changes in the measured current or voltage have to be processed. Hence, the sensitivity and stability of the input amplifier must be very high in order to detect such small changes in the measured signal.

Independent from the measurement method, careful consideration of measurement errors is necessary. A main source of measurement errors may be parasitic components, such as stray capacitances between neighboring wires leading to the sensing electrodes, or between wires and their shielding, or stray capacitances between metallic components of the measuring system and ground, which become the more effective the higher the measuring frequency.

The risk for undesired stimulation of the heart or peripheral nerves if such electrical voltages or currents are applied for monitoring purposes, is negligible, both with regard to the high frequency and the low current density. Furthermore, heating and heat-induced secondary effects can be neglected.

However, proper attention must be paid for the selection of the equipment and its performance data for the intended application. Furthermore, the employed devices must be safe even in case of technical failure. Patient-near devices are directly connected with the patient whereby the connecting impedance is rather low.

CHARACTERISTICS OF BIOIMPEDANCE

The microscopic electrical properties that describe the interaction of an electromagnetic wave with biological tissue are the complex conductivity σ^* with the unit $\Omega^{-1}{\cdot}\mathrm{m}^{-1}$, mho$\cdot$m^{-1}, S$\cdot$m^{-1}, or $1{\cdot}(\Omega{\cdot}\mathrm{m})^{-1}$

$$\sigma^*(\omega) = \sigma'(\omega) + j\sigma''(\omega)$$

and the complex dielectric permittivity ε^* with the unit F/m

$$\varepsilon^*(\omega) = \varepsilon'(\omega) - j\varepsilon''(\omega)$$

ω is radian frequency with the unit hertz. The electrical properties are depending on the frequency with strong dependence in the range of a dispersion.

The relation between these two quantities can be described in accordance with Ref. 13 by

$$\sigma^*(\omega) = j\omega\,\varepsilon^*(\omega)$$

With the conduction current that is related with the basic conductivity σ_0, that is, the current carried by the mobility of ions in the extracellular space, and the polarization current (sometimes called displacement current) that is related with permittivity, the following equations are obtained

$$\sigma' = \sigma_0 + \omega\,\varepsilon''(\omega)$$
$$\sigma'' = \omega\,\varepsilon'(\omega) = \varepsilon_0\,\varepsilon_r(\omega)$$

where ε_0 is the dielectric permittivity of free space with $\varepsilon_0 = 8.85 \times 10^{-12}$ F·m^{-1}, and ε_r is the relative dielectric permittivity (with $\varepsilon_r = 1$ for the free space and $\varepsilon_r = 81$ for water in the low and medium frequency range).

Instead of the complex conductivity σ^*, the inverse complex resistivity ρ^* with the unit Ω·m can be used. The resistivity is usually preferred in the context of impedance plethysmography:

$$\rho^*(\omega) = \rho'(\omega) + j\,\rho''(\omega)$$

The complexity of these quantities considers the fact that in the alternating current (ac) range the biological tissue cannot adequately be described by a simple resistance (or its inverse conductance), but needs the extension to a complex quantity, that is, impedance or admittance. Some authors prefer the term Admittance Plethysmography instead of Impedance Plethysmography (14,15). The simplest adequate model for such an impedance is represented by a resistance and a reactance. The resistance causes the loss in power, whereas the reactance causes the delay (or 18 phase shift) between voltage and current. The dominating reactance of bioimpedance in the frequency range of interest is capacitive and becomes more relevant for higher frequencies. Only for very high frequencies that usually are not employed for impedance plethysmography, can the reactance be composed by both a capacitive and an inductive component.

Bioimpedance can be described like any technical impedance in different forms, for example, by its magnitude (or modulus) Z_0 and its phase angle (or argument) φ, (i.e., the delay between voltage and current):

$$Z = Z_0\,e^{j\varphi}$$

or by its real part (or resistance) Re{Z} and its imaginary part (or reactance) Im{Z}:

$$Z = \mathrm{Re}\{Z\} + j\,\mathrm{Im}\{Z\}$$

Alternating current voltages and ac currents, too, can be expressed as complex quantities, that is, by their magnitude and phase angle, or by their real and imaginary part although this is rather unusual. The magnitude of the impedance Z_0 corresponds to the quotient of the magnitudes of voltage V_0 and current I_0, that is,

$$Z_0 = V_0/I_0$$

Appropriate modeling of the electrical properties of biological tissue by discrete and lumped electrical components renders possible the proper consideration of multilayer or compartmentally composed tissues with different electrical properties of each layer or compartment. Such tissues can be modeled as serial, parallel, or serial–parallel equivalent circuits in 2D presentation. More recently, the modeling has been extended to 3D models using the FEM or comparable approaches.

The impedance parameters can be depicted in different modes as a function of frequency (Fig. 4). The presentation of the magnitude (usually in logarithmic scale with regard to its wide range) and the phase angle against the frequency over several decades, and therefore in logarithmic scale is known as the Bode plot. A similar presentation is used for both the real and imaginary part versus the frequency on the x axis. This mode of presentation is sometimes called the spectrum (e.g., modulus spectrum and phase angle spectrum). A different form of presentation is in a plane with the real part along the x axis and the imaginary part along the y axis, both in linear scaling, with the frequency as parameter. This mode of presentation is frequently called the Cole–Cole-plot (but also the Nyquist plot, locus plot, or Wessel graph). The same modes of

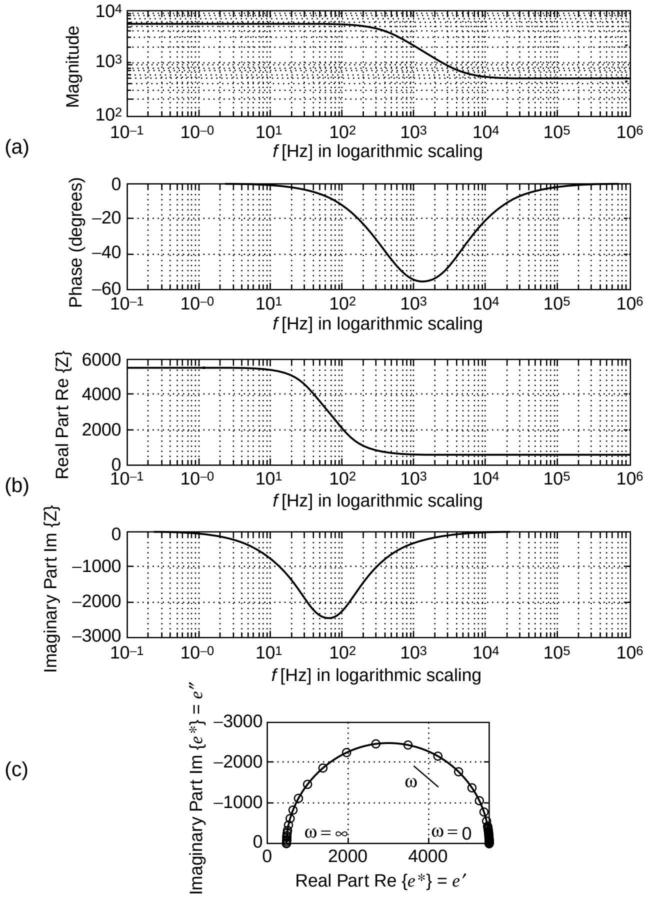


Figure 4. Different modes for the presentation of impedance quantities. (a) Shows the magnitude (in logarithmic scaling) and the phase angle of impedance as functions of frequency in logarithmic scaleing (Bode plot). (b) Shows the real and imaginary part of impedance depicted versus the frequency in logarithmic scaling. (c) Presents the Cole–Cole plot with the real part at the x axis, the imaginary part at the y axis, and the frequency as parameter along the curve. The results shown are not from biological tissue, but calculated for the circuit of Fig. 1b with $R_s = 500$ Ω, $R_p = 5$ kΩ, and $C_p = 500$ nF.

presentation are possible for the complex quantities σ^*, ρ^*, and ε^*.

Usually, impedance plethysmography is accomplished employing only a single measuring frequency or a few discrete measuring frequencies. However, impedance spectroscopy with a multitude of measuring frequencies is gaining interest, especially for the determination of spatially distributed volumes. Typical examples are the determination of body composition, tissue, or organ vitality monitoring in combination with cellular edema as a result of hypoxemia, and monitoring of infusion therapy.

Certain forms of electrotherapy are also utilizing the passive electrical properties of biological tissue. Methodology and technology for these forms of electrotherapy, however, are not discussed here.

MODEL-BASED RELATIONS FOR VOLUME DETERMINATION

Valid relations between the monitored electrical quantities and the searched volumetric parameters have to be used to calculate a quantitative result that can be expressed in units of volume (e.g., mL or cm^3). Most of these relations are based on models. Possibly the first model for the interpretation of bioimpedance measurements has already been developed for a suspension of cells in a fluid by Fricke and Morse in 1925 (16).

The simplest model-based approach for establishing an impedance–volume relationship is a cylindrical volume conductor of radius r_0, length L, and resistivity ρ^*. It is assumed that this volume conductor is surrounded by soft material with significantly higher resistivity $\rho_e^* \gg \rho^*$, so that it must not be considered as a parallel circuit and its actual radial extension has no impact. This volume conductor may be a blood vessel (e.g., an artery or vein) surrounded by tissue that has higher resistivity than blood. Furthermore, it is assumed that the inflow of the volume ΔV into that cylinder expands the radius homogeneously by Δr over the total length. It is also assumed that neither the length L nor the resistivity ρ^* are affected by the volume injected into the volume conductor. For didactic simplicity, only the real part ρ' of the complex resistivity ρ^* is considered; that is, only the real part of the impedance is taken into account. However, despite this simplification, the variable is understood as impedance Z. This simplification is generally valid for frequencies much lower than the β-dispersion, since for these frequencies the phase angle is small ($< 10^\circ$). For higher frequencies, the calculation must be performed with proper consideration of the complex quantities.

With these assumptions the following relations for Z and V are valid:

$$Z_0 = \rho' L/[\pi r_0^2]$$
$$V_0 = L\pi r_0^2$$

From these two equations the following equation can easily be calculated

$$Z_0 = \rho' L^2/V_0$$

After the inflow of the volume ΔV and the increase of the radius by Δr, the following relations are valid:

$$Z_1 = Z_0 - \Delta Z = \rho' L/[\pi(r_0 + \Delta r)^2]$$
$$V_1 = V_0 + \Delta V = L\pi(r_0 + \Delta r)^2$$

with $Z_1 < Z_0$ and $V_1 > V_0$ for $\Delta r > 0$.

With some simple mathematical operations it can be shown that

$$\rho' L^2 Z = (\rho' L^2 + \Delta V Z)(Z - \Delta Z)$$

If the product $\Delta V\, \Delta Z\, Z$ is neglected as a product of small quantities, the result becomes:

$$\Delta V = \rho'(L/Z)^2\, \Delta Z$$

This is the well-known Nyboer equation that relates the volume change ΔV with the change in impedance ΔZ as a consequence of the blood inflow into a peripheral artery, for example, into the aorta or carotid artery with each heart beat.

For proper application all included simplifications must carefully be taken into account. Only mathematical simplifications, but no methodological constraints have been mentioned here. Such a constraint may be that a part of the injected volume is already flowing out of the measured vessel segment before the inflow of the whole volume ΔV into this vessel segment has been completed. This constraint must especially be considered if the segment is short and the vessel wall rather stiff.

With regard to the surrounding tissue, a more realistic model would be an arrangement of two concentric cylinders of length L with different conductivities. The inner cylinder with radius r_1 and resistivity ρ_1' has the impedance $Z_1 = \rho_1' L/[\pi r_1^2]$, whereas the outer cylinder with radius r_2 and resistivity ρ_2' has the impedance $Z_2 = \rho_2' L/[\pi(r_2 - r_1)^2]$. Electrically both cylinders are arranged in parallel configuration. Hence, the total impedance is obtained by $Z_0 = Z_1 Z_2/(Z_1 + Z_2)$. The inner cylinder shall be a vessel into which blood of volume ΔV is pumped, causing a homogenous dilation of the vessel radius by Δr_1 and a lowering of its impedance to $Z_1^* = Z_1 - \Delta Z_1 = \rho_1' L/[\pi(r_1 + \Delta r_1)^2]$. Since Z_2 shall not be affected, the total impedance is $Z_0^* = Z_1^* Z_2/(Z_1^* + Z_2)$. The following steps are similar to those leading to the Nyboer equation. Since the resulting equation and its application to measurements are more complicated, they will not be discussed here in detail. Even this model is actually simplified, because in reality the tissue around the vessel will neither be a cylinder nor have a homogeneous resistivity. This last situation may become relevant with a vein or a bone in the vicinity of the artery.

With regard to the constraints of the Nyboer equation, another approach has been used that finally leads to the Kubicek equation (17). The model-based approach starts again with the single-vessel model of length L. However, in contrast to the Nyboer approach the assumption is not made that the inflow of the volume ΔV into the considered vessel segment is finished before the outflow starts. Here, the basic assumption is that the inflow is constant during the inflow time T_{inf} and that the outflow starts with some delay, however, temporal overlap of outflow with inflow must not be excluded. With this assumption, the change in the intravascular volume and, hence, in the impedance, is maximal when there is only inflow into and no outflow from the segment. This maximal change of the impedance can be expressed by its first time derivative [i.e., by $(dZ/dt)_{\max}$]. The total inflowing volume ΔV can then be taken into account by multiplying $(dZ/dt)_{\max}$ with the inflow time T_{inf}. With regard to the aorta this inflow time is equivalent with the ejection time of the left ventricle. In many cases even the inflow time can additionally be obtained from the impedance curve. This leads finally to the Kubicek equation:

$$\Delta V = \rho'(L/Z)^2 T_{\text{inf}} (dZ/dt)_{\max}$$

Obviously, the only relevant difference in both approaches is the Nyboer assumption that the total volume change ΔV has already been injected into the measured vessel segment before the outflow starts against the Kubicek assumption that this volume ΔV is entering the measured vessel segment with constant rate during the whole inflow period. The Kubicek equation is more realistic for a short vessel segment with a rather stiff wall. For such vessels, the Nyboer equation leads to an underestimation of the real volume change. In contrast, if the inflow is decreasing at the end of the inflow period, for example, at the end of the ventricular ejection period, the Kubicek equation yields an overestimation of the volume change.

All other model-based assumptions are identical or comparable. Both approaches consider only a single vessel with homogeneous dilation over the total length within the measured tissue and neglect the surrounding tissue and its composition with regard to nonhomogeneous resistivity. Blood resistivity is taken as constant although there is some evidence that it depends on the flow velocity.

Although the Kubicek equation has primarily been proposed for the monitoring of cardiac output, both equations have also been applied to the monitoring of pulsatile peripheral blood flow. Both models, however, do not consider that in the peripheral circulation a basic or nonpulsatile blood flow may exist as well.

Different modifications have been suggested in order to overcome relevant drawbacks. Most of these modifications are optimized with regard to the monitored quantity, geometric constraints, modes of application, or positioning and shape of electrodes. They will not be discussed here.

No valid impedance–volume models have been proposed for the quantitative monitoring of ventilation by the application of impedance plethysmography. Statistical models are used for the impedance–volume relationship regarding body composition. Some first approaches have been suggested for the volume changes due to fluid shifts.

INSTRUMENTATION AND APPLICATIONS

The typical basic equipment for impedance plethysmography consists of the signal generator, either a constant voltage generator or a constant current generator; the frequency-selective measuring device, either for current or voltage, in combination with AD conversion. The equipment may be supplied with more than one signal channel for certain applications, for example, with two channels for simultaneous and symmetric monitoring at both extremities or one channel for each frequency in multifrequency measurements; the signal processor, for example, for processing the impedance quantities; the processing unit for calculating the volumetric quantities; the monitor and/or data recorder; multiple electrodes and shielded leads; specific auxiliary equipment, for example, venous occlusion machine with cuff and pump.

Devices for impedance plethysmography are small, light, usually portable, and battery powered. The devices for patient-near application are much cheaper than competitive equipment based on nuclear magnetic resonance (NMR), X ray, or US. Also, the running costs are much lower than for the competitive technologies, usually these costs are mainly required by the single-use electrodes.

Peripheral Hemodynamics

The objective is the detection of deficiencies either in the arterial or venous peripheral circulation. The application of impedance plethysmography to peripheral vascular studies has already been in the interest of Nyboer in 1950 (18).

In the peripheral circulation, the most interesting quantity is arterial blood flow or perfusion. Impedance measurement is performed either in the bipolar or, more frequently, in the tetrapolar configuration. The tetrapolar configuration requires a longer segment for measurement in order to place the sensing electrodes in proper distance from the feeding electrodes with the nonhomogenous current field in their vicinity. Electrodes are either of the circular or disk or the band type. Disk electrodes can be placed directly above the monitored vessel and therefore provide high sensitivity, but the magnitude and the reproducibility of the measured signal in repeated measurements are strongly dependent on exact electrode placement (19–21). Band electrodes are preferred for the measurements at extremities because they can be placed around the extremities. In this case, the measured object is the whole segment between the sensing electrodes including the extravascular tissue and may include more than only one vessel. Flow can be estimated by application of the Nyboer, the Kubicek or any modified impedance–volume equation. Competitive methods are utilizing ultrasound Doppler, contrast X-ray angiography, or NMR.

Some diagnostic information about the stiffness of the arterial vessel wall can be obtained by the impedance method from the measurement of the pulse wave propagation velocity, usually executed at two different sites of the same arterial pathway. The pulse that is actually recorded with the impedance method is the intravascular volume pulse, that is, the dilation of the vessel with each heart beat (22). Simple formalistic models are used to relate the pulse wave propagation velocity with the stiffness of the vessel wall. If the intravascular blood pressure is also monitored, then it is possible to calculate the stiffness or its inverse, the compliance as ratio of the volume change and pressure change $\Delta V/\Delta p$, directly.

Another problem is the diagnosis of proximal or deep venous thrombosis and of other obstacles for the venous return flow to the heart from the extremities (23). One approach is actually a modification of the venous occlusion plethysmography that has already been introduced in 1905 by Brodie and Russel (24). A cuff is placed around the extremity and connected with a pump. The cuff pressure is enhanced abruptly so that it occludes the vein and stops venous outflow without affecting the arterial inflow. The volume increase following venous occlusion allows estimating the arterial inflow. When the occlusion is stopped after $\sim$ 20 s, the venous outflow starts again and thereby leads to a reduction in volume. The slope or the time constant of this postocclusion emptying process are used to assess the outflow resistance, for example, the hydrodynamically obstructive impact of deep venous thrombosis, or the venous wall tension. However, other pathological effects must be carefully considered, for example, increased central venous pressure. For this reason, the recording is frequently and simultaneously performed on both extremities, so that the results can be compared with each other. The measurement is usually executed with band electrodes. Competitive methods are ultrasound Doppler, contrast X-ray venography, or NMR.

A similar impedance-based approach is employed to test the performance of the drainage system in extremities. Changes in the hydrostatic pressure are used to shift volume between the trunk and an extremity, for example first by bringing down an arm before raising it above the head. The affected volume shifts can be recorded and render possible the assessment of the performance of the draining system. This approach is frequently used to study fluid shifts caused by tilting experiments, during microgravity experiments, or after long periods of bedrest.

Brain Perfusion and Edema

The most important objectives are monitoring of cerebral bloodflow and the detection of cerebral edema. First publications about rheoencephalography are dating back to 1965 (25,26).

The volume of the brain with its enclosed fluid spaces, for example, the intravascular volume and the cerebrospinal fluid volume, is kept constant by its encapsulation in the bony skull. The expansion of the volume of one compartment, for example, increase of the intravascular volume by augmented arterial blood pressure, the space-demanding growth of a brain tumor or intracerebral bleeding, can only be compensated by the diminution of the volume of other compartments. If the space-demanding process is of nonvascular origin, the most affected compartment will be the intravascular volume. Due to the compression of blood vessels, the cerebral bloodflow and thus metabolism will be reduced.

Impedance measurements aiming for the brain as organ are difficult because the encapsulating bony skull has a very high resistivity as compared with the soft extracranial tissue of the face and the scalp. If the tetrapolar mode is used, more than two sensing electrodes may be applied. Different electrode arrangements have been described to force the current pathways through the skull into the brain, but also the application of the feeding electrodes to the closed eyelids. However, the measurement of the transcephalic impedance has not become a routinely applied clinical method with the exception of neonates in which the thickness of the bony skull is very small. Competitive methods based on NMR, X ray, US, and even photoplethysmography have gained more attention in the recent past. Some expectations are related to the development of contactless applications, especially for the continuous monitoring of edema (27). This might allow control treatment by hyperosmolaric infusion therapy.

Ventilation and Lung Performance

Impedance pneumography were among the first applications of impedance plethysmography and had already been described by Geddes et al. in 1962 (28,29).

The objective of impedance pneumography is to record the tidal volume under resting conditions or during exercise. Additionally, the breathing rate can be obtained by the impedance method. The principle is based on the measurement of the transthoracic impedance that increases during inspiration as a consequence of increasing alveolar air filling, and decreases during expiration (30). The conductivity of lung tissue at the end of a normal expiration is $\sim$0.10 $\Omega^{-1}\cdot\mathrm{m}^{-1}$ as compared with 0.05 $\Omega^{-1}\cdot\mathrm{m}^{-1}$ at the end of normal inspiration. The application of impedance pneumography is very simple and also applicable for critically ill patients, since it allows continuous recording without requiring a breathing tube. However, the quantitative determination of the tidal volume is difficult and needs calibration by another spirometric method. No realistic model-based interpretation of quantitative assessment is available until now. Some expectations are related with multifrequency measurement (31).

The impedance measurement can be performed with the bi- or tetrapolar configuration. It is usually performed separately for each side of the lung in order to detect differences. Even with more electrodes, however, the spatial resolution is too poor to allow detection of regional inhomogeneities of alveolar air filling. For that reason, this field is gaining growing interest for the application of EIT (4). Also, EIT has limited spatial resolution, as compared with other CT-based imaging procedures. Despite this drawback, it has some potential for the detection of regional inhomogeneities in ventilation. Such an approach would be of high relevance for diagnostic purposes and for the efficiency control of artificial ventilation. Serious drawbacks for EIT, however, are the costs of the equipment and the necessity to attach up to 64 or 128 electrodes around the thorax.

Since pulmonary edema is usually a general and not a localized phenomenon, its monitoring might be possible by utilizing the transthoracic impedance measurement. Measurement of extravascular lung water is investigated as a methodological approach to guide the fluid management of patients with noncardiogenic pulmonary edema (32). The conductivity of lung edema fluid is $\sim$1 $\Omega^{-1}\cdot\mathrm{m}^{-1}$, and therefore is distinctly different from alveolar tissue filled with more or less air.

The impedance coefficients of tumor tissue are different from normal lung tissue. Hence, cancer detection might be possible by bioimpedance measurement. But with regard to its poor spatial resolution, the bi- or tetrapolar transthoracic impedance measurement is not qualified, but EIT might become useful for some special applications. Other imaging procedures were superior until now, primarily due to the higher spatial resolution as compared with EIT.

Much work has been devoted to comparing impedance pneumography with inductive pneumography. There is some evidence that inductive pneumography is superior concerning ventilation monitoring in newborn infants, especially for risk surveillance with regard to SIDS. An interesting approach tries to utilize inductive pneumography in combination with the evaluation of the signal morphology for the monitoring of airway obstruction (33).

Intercompartmental Fluid Shifts

Intercompartmental fluid shifts occur during dialysis, for example, hemodialysis, but also during infusion therapy and emergence of edema. The assessment of fluid shifts, which are actually changes of volumes, by the measurement of electric variables has been an outstanding objective very early in the scientific research and medical utilization of bioimpedance and dates back to 1951 (34).

Hemodialysis is a therapeutic procedure that is employed in patients with renal insufficiency. The therapeutic objective is to remove water, electrolytes, urea, and other water-soluble substances in combination with the reestablishment of a normal acid–base status. In a simplified model, the water is first removed from the intravascular volume, (i.e., the blood plasma). This means that the hematocrit, and thereby the viscosity of blood, is increased cousing the work load for the heart is enhanced. Also, the osmotic pressure of the blood is raised, whereas the hydrostatic blood pressure is lowered. Both effects contribute to the refilling of the intravascular space by a fluid shift from the interstitial space (i.e., the extravascular extracellular space). This fluid shift finally causes another fluid shift from the intracellular space into the interstitial space. The dynamics of these fluid shifts is primarily controlled by the hydrostatic and osmotic pressure differences between the different spaces, but also by the substance-specific permeability of the different barriers between the spaces including the dialysis membrane. If removal of water is too fast or changes of ions like sodium, potassium, and calcium are too large, the hemodynamics of the patient or the excitability of tissues like the heart or central nervous system may become disturbed. Impedance plethysmographic methods have some potential for the control of those fluid shifts, that is, may help to avoid critical disequilibrium syndromes like hypotension, headache, and vomiting. The best results of the plethysmographic measurement are achieved if segmental measurement is performed at the extremities instead of whole body measurements (5). Figure 5 shows a schematic presentation of such segmented body. The forearm accounts only for ~1% of body weight, but contributes 25% to whole body impedance.

Infusion therapy aims mainly to filling the intravascular volume by utilizing venous access. However, depending on the control variables (e.g., hydrostatic pressure, osmotic pressure, and permeability of the barriers), intercompartmental fluid shift cannot be avoided. Consequently, part of the infused volume will not remain in the intravascular system and help to stabilize blood pressure, but escape to the extravascular space.

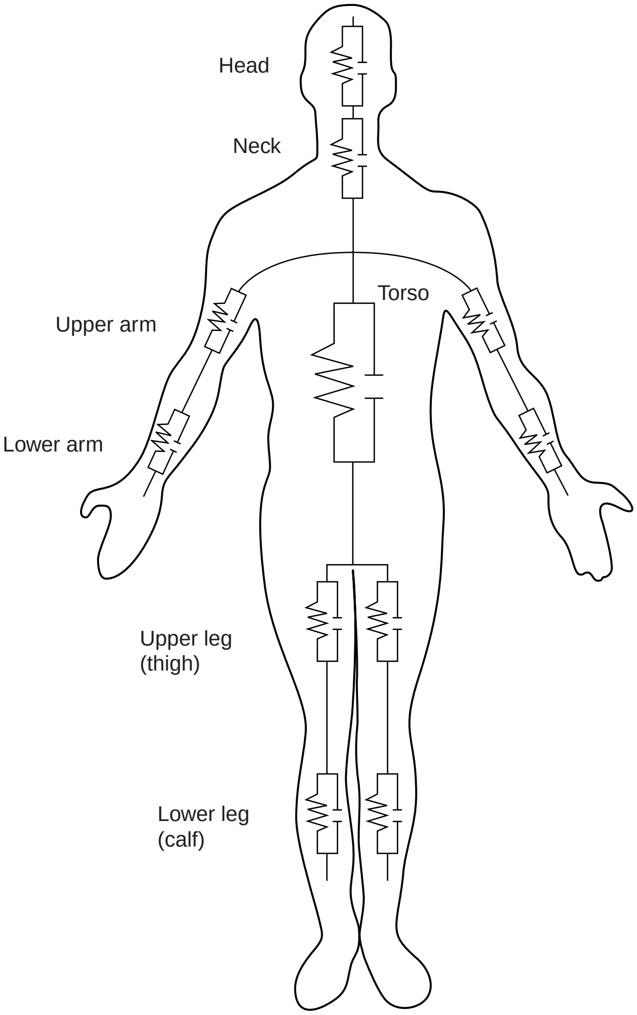


Figure 5. Schematic presentation of the whole body subdivided into 11 segments that are modeled by *RC*-networks.

A special objective of intravenous infusion therapy with hyperosmolaric fluid is the removal of fluid from the extravascular space in order to avoid the emergence of edemas. In the brain, such infusion may help to lower the intracranial pressure that is caused by edema.

Body Composition

Most frequently this method is called bioelectric impedance analysis (BIA). The interest in this methodological approach has increased with the increased interest in healthy life style in industrialized populations since the first reports at the end of the 1980s (35,36).

The body consists of different tissues, for example, muscle, fat, bones, skin, the nervous system, and connective tissue. All these tissues contain intra- and extravascular spaces. The extravascular space can be subdivided into the extracellular or interstitial space and the intracellular space. Chemically, the body consists of water, proteins, carbohydrates, lipids, ions, rare elements, and some other substances. The main objective of body composition analysis is the assessment of total body water, of lean or fat free tissue mass, and of fat tissue. Those measurements have clinical relevance, but they are also gaining growing attention in the field of sports, physical fitness, wellness, nutrition, and for life science in aerospace research.

The methodological approaches are different, for example bipolar or tetrapolar mode, single-or multiple-frequency measurement, whole body or segmental measurement. In single-frequency measurements, a sinussoidal current with usually < 1 mA (range 0.02–10 mA) and a frequency with typically 50 kHz (range 20–100 kHz) is employed. The basic assumption is that resistivity of fat tissue is higher than that of the so-called lean tissue that has a higher content of water. In the nonclinical field, the determination of body composition is frequently combined with the measurement of body weight using a scale with two integrated electrodes in the foot platform. In that case, however, only the two legs and the lower part of the trunk will be included into the estimation. In more advanced devices, additional electrodes are available in the form of hand grips. Detailed segmental measurements are more valid since the trunk of the body may contribute $\sim$70% to the total body weight and up to 90% to the total body fat, but only 5% to the measured whole body impedance. More sophisticated equipment utilize multifrequency measurement (11).

The applied statistically based predictive equations are of the linear regression type and consider individual parameters like sex, age, height, and weight. They are primarily used to assess the lean muscle mass, the body fat, the fat free mass, the water content, and the body mass index. For physically based segmental calculations, the extremities, the body trunk, and even the head are modeled by cyclindric shape with uniform cross-section over the total length.

Competitive methods include antropometric measures like the girth, simple skin-fold measurements by mechanical calipers, but also highly advanced methods, such as NMR, X ray (dual energy X-ray absorptiometry, or nuclear imaging), and for certain purposes even the hydrostatic weighing in a water tank that is assumed to be the most accurate method.

Laboratory Applications

Blood Cell Counting. The employed methodological principle is frequently called the Coulter principle (37). Blood cells (e.g., erythrocytes, leucocytes) are passing through a capillary (diameter < 100 μm) filled with blood plasma. For frequencies below the β-dispersion the impedance of the cell is higher than that of the surrounding plasma. Consequently, the passage of each cell affects the recorded impedance. Those impedance changes are used for cell counting. In sophisticated devices the impedance changes are quantitatively measured and allow cell volume estimation, which also renders possible the determination of the cell volume distribution function (frequently called the Price–Jones distribution function with the cell diameter as variable). Since the number of leucocytes is very small compared with that of the erythrocytes (usually $< 0.1\%$), leucocytes do not really disturb the counting of erythrocytes. In contrast, however, the erythrocytes must be destroyed before the leucocytes can be counted. This is usually achieved by chemical hemolysis of the erythrocytes. Furthermore, chemical substances are utilized to render possible the differentiation between the different populations of leucocytes (e.g., granulocytes, monocytes, lymphocytes).

Hematocrit. The objective is the estimation of the intracellular volume of blood. The hematocrit is defined as the ratio of the volume of blood cells to total blood volume, although frequently the hematocrit is taken as a measure for the ratio of only the volume of erythrocytes to total blood volume. However, since the partial volume of all other blood cells (i.e., leucocytes, platelets) is very small compared with that of the erythrocytes, the results are not distinctly different. Determination of the ratio between the extracellular volume and the total blood volume is possible by application of at least one measuring frequency below and another one above the β-disperion. Problems that need further consideration as possible sources of error are the electrolytic conductivity of blood plasma, which is dependent on the actual concentration of ions, and the sedimentation of the cells as a consequence of their higher specific weight that may take place during the measuring period. The measurement requires the withdrawal of a sample of blood from a vein or another blood vessel. Measurement of the hematocrit can be achieved either with the impedance or the dielectric technique (38).

Several modifications of the impedance method for the noninvasive determination of the hematocrit have been proposed. In one approach, impedance measurement is performed at the finger by application of two different frequencies (i.e., 100 kHz and 10 MHz). The hematocrit is determined by an algorithm that uses both the pulsatile and the baseline component of both measuring frequencies (39). In another remarkable methodological approach, the patient puts a finger in a temperature-stabilized bath. A fluid with gradually increasing ionic concentration is pumped through the bath chamber, thus leading to a decrease in the impedance of the bath fluid. The pulsatile volume changes of the finger are recorded by impedance measurement in the bath chamber. These pulsatile impedance fluctuations disappear only if the impedance of the bath fluid is identical with that of the blood in the finger. The conclusion is made from the actual impedance of the bath fluid on the hematocrit (40).

Others

Cell Imaging. Nanotechnology-based on bioimpedance sensing, chip devices have been described that allows us to rapidly detect and image cells with a specific phenotype in a heterogeneous population of cells (41). This might be useful for screening purposes, recognition of cell irregularities, and detection of risk patients like human immunodeficiency virus (HIV)-infected individuals. Cell identification is made possible by administration of marker substances. A promising measurement procedure is electrochemical cyclic voltammetry. When a sinusoidal voltage with constant amplitude and a variable frequency in the range of some kilohertz is applied, the impedance is plotted in the spectrographic mode. Among other effects, volume changes might be the dominating measured effect if the marker binds with a receptor in the cell membrane, and herewith affects membrane properties like its permeability for water or ions or the transmembraneous active ion transport mechanisms.

Inductive Plethysmography. Inductive plethysmography is employed for respiratory monitoring, so-called respiratory inductance plethysmography (RIP). It is based on the measurement of the thoracic crosssection that is enclosed by coils and includes both the rib cage and abdominal compartments (42–45). In the medium frequency range (30–500 kHz), changes in the volume are monitored by the influenced inductance. In the higher frequency range (~100 MHz), the inductively provoked signals (i.e., eddy currents depending on the alveolar air filling) are recorded by appropriately arranged coils.

Magnetic Susceptibility Plethysmography. Magnetic susceptibility plethysmography is a contactless method. It is based on the application of a strong magnetic field and monitors the variation of the magnetic flux. The measurement is accomplished with superconduting quantum interference device (SQUID) magnetometers. This approach may primarily be utilized for the assessment of blood volume changes in the thorax, but until now it is not employed for clinical routine (46).

SUMMARY

In comparison with biomedical engineering as a recognized discipline, the research activities in the field of bioimpedance are much older. It can be assumed that Nikola Tesla, a former student of physics in Graz (Austria) and the inventor of the ac technology, already knew about the passive electrical properties of biological tissues when he demonstrated his famous and public performances with the administration of high voltage pulses "···I demonstrated that powerful electrical discharges of several hundred thousand volts which at that time were considered absolutely deadly, could be passed through the body without inconvenience or hurtful consequences" in the 1880s. This knowledge was utilized by d'Arsonval since 1892 for therapeutic purposes, mainly aiming for heat induction in certain parts of the body. In 1913, Rudolf Hoerber, at that time a physiologist at the University of Kiel (Germany), measured the electrical conductance of frog muscle at 7 MHz and found that at this frequency the membrane resistance is short circuited.

Since its beginning, bioimpedance remained to be a challenge to physicists, medical doctors, and of course engineers. The most relevant basic research was performed in the second half of the twentieth century. The progress that has been reached has been and is still utilized both for diagnostic and therapeutic purposes in medicine. Impedance plethysmography is one of the different fields of bioimpedance application. If impedance plethysmography is correctly understood, it does not only mean the determination of a solid volume with well-defined boundaries, but also the volumetric determination of one component contained in a mixture of different components.

Progress in technology has rendered possible applications that are of great interest for medicine. The most relevant progress is in the field of signal acquisition, including advanced electrode technology, signal processing, and model-based signal interpretation. Not all attempts to utilize the passive electrical properties of biological tissue for diagnostic purposes have been successful. In many cases, other technologies have been shown to be superior. But there is no doubt that the whole potential of impedance plethysmography has not been exhausted. New challenges in the medical field are cellular imaging and possibly even molecular imaging. In all applications, however, impedance plethysmography will have to prove its validity and efficiency.

BIBLIOGRAPHY

1. Nyboer J, Bango S, Nims LF. The Impedance Plethysmograph and Electrical Volume Recorder. CAM Report OSPR 1943; 149.
2. Schwan HP. Electrical properties of body tissues and impedance plethysmography. IRE Trans Biomed Electron 1955; 3:32–46.
3. Stewart GN. Researches on the circulation time in organs and on the influence which affect it. J Physiol 1894;15:1–89.
4. Li J. Multifrequente Impedanztomographie zur Darstellung der elektrischen Impedanzverteilung im menschlichen Thorax. PhD thesis, University of Stuttgart, 2000. Available at http://elib.uni-stuttgart.de/opus/volltexte/2000/736/pdf/li_diss.pdf.
5. Kanai H, Haeno M, Sakamoto K. Electrical measurement of fluid distribution in legs and arms. Med Prog Technol 1987;12:159–170.
6. Osten H. Impedanz-Plethysmographie im Orthostasetest. Münchn Med Wochenschr 1977;119:897–900.
7. Gersing E. Measurement of electrical impedance in organs. Biomed Techn 1991;36:6–11.
8. Dzwonczyk R, et al. Myocardial electrical impedance responds to ischemia in humans. IEEE Trans Biomed Eng 2004;BME-51:2206–2209.
9. Durney CH, Massoudi H, Iskander MF. Radiofrequency Radiation Dosimetry. 4th ed. USAFSAM-TR-85–73; 1985.
10. Gabriel S. 1997. Appendix B: Part 1: Literature Survey. Available at http://niremf.ifac.cnr.it/docs/DIELECTRIC/AppendixB1.html.
11. Segal KR. Estimation of extracellular and total body water by multiple-frequency bioelectrical impedance measurement. Am J Clin Nutr 1991;V54-1:26–29.
12. Neuman MR. Biopotential electrodes. In: Webster JG, editor. Medical Instrumentation—Application and Design. 3rd ed. New York: Wiley; 1998.
13. Rigaud B, Morucci J-P, Chauveau N. Bioimpedance measurement. In: Morucci J-P, et al. edition. Bioelectrical Impedance Techniques in Medicine. Crit. Rev. Biomed. Eng. 1996; 24: 257–351.
14. Yamakoshi KI, Shimazu H, Togawa T, Ito H. Admittance plethysmography for accurate measurement of human limb blood flow. Am J Physiol 1978;235:H821–H829.
15. Shimazu H, et al. Evaluation of the parallel conductor theory for measuring human limb blood flow by electrical admittance plethysmography. IEEE Trans Biomed Eng 1982;BME-29: 1–7.
16. Fricke H, Morse S. The electrical resistance of blood between 800 and 4.5 million cyclces. J Gen Physiol 1925;9: 153–157.

17. Kubicek WG, et al. Development and evaluation of an impedance cardiac output system. Aerospace Med 1966;37:1208–1212.
18. Nyboer J. Electrical impedance plethysmography: a physical and physiological aproach to peripheral vascular studies. Circulation 1950;2:811–821.
19. Yamamoto Y, Yamamoto T, Öberg PA. Impedance plethysmography in human limbs. Part 1: On electrodes and electrode geometry. Med Biol Eng Comput 1991;29:419–424.
20. Yamamoto Y, Yamamoto T, Öberg PA. Impedance plethysmography in human limbs. Part 2: Influence of limb crosssectional areas. Med Biol Eng Comput 1992;30:518–524.
21. Lozano A, Rosell J, Pallas-Areny R. Errors in prolonged electrical impedance measurement due to electrode repositioning and postural changes. Physiol Meas 1995;16:121–130.
22. Risacher F, et al. Impedance plethysmography for the evaluation of pulse wave velocity in limbs. Med Biol Eng Comput 1992;31:318–322.
23. Hull R, et al. Impedance plethysmography: The relationship between venous filling and sensitivity and specificity for proximal vein thrombosis. Circulation 1978;58:898–902.
24. Brodie TG, Russel AE. On the determination of the rate of blood flow through an organ. J Physiol 1905;33:XLVII–XLVIII.
25. Seipel JH, Ziemnowicz SAR, O'Doherty DS. Cranial impedance plethysmography—rheoencephalography as a method for detection of cerebrovascular disease. In: Simonson E, McGavack TH, editors. Cerebral Ischemia. Springfield, IL: Charles C Thomas; 1965; p 162–179.
26. Hadjiev D. A new method for quantitative evaluation of cerebral blood flow by rheoencephalography. Brain Res 1968;8:213–215.
27. Netz J, Forner E, Haagemann S. Contactless impedance measurement by magnetic induction—a possible method for investigation of brain impedance. Physiol Meas 1993;14:463–471.
28. Geddes LA, Hoff HE, Hickman DM, Morre AG. The impedance pneumograph. Aerospace Med 1962;33:28–33.
29. Baker LE, Geddes LA, Hoff HE. Quantitative evaluation of impedance spirometry in man. Am J Med Elect 1965;4:73–77.
30. Nopp P, et al. Dielectric properties of lung tissue as a function of air content. Phys Med Biol 1993;38:699–716.
31. Brown BH, et al. Multifrequency imaging and modelling of respiratory related electrical impedance changes. Physiol Meas 1994;15(Suppl. A):1–12.
32. Nierman DM, et al. Transthoracic bioimpedance can measure extravascular water in acute lung injury. J Surg Res 1996;65:101–108.
33. Brack Th et al. Continuous and cooperation-independent monitoring of airway obstruction by a portable inductive plethysmograph. AJRCCM 2004;169:1.
34. Löfgren B. The electrical impedance of a complex tissue and its relation to changes in volume and fluid distribution. Acta Physiol Scand 1951;23(Suppl. 81):1–51.
35. Schloerb PR, et al. Bioimpedance as a measure of total body water and body cell in surgical nutrition. Eur Surg Res 1986;18:1.
36. van Loan MD, et al. Association of bioelectric resistive impedance with fat-free mass and total body water estimates of body composition. Amer J Human Biol 1990;2:219–226.
37. Coulter WH. High speed automatic blood cell counter and cell size analyzer. Proc Nat Electron Conf 1956;12:1034.
38. Treo EF, et al. Comparative analysis of hematocrit measurements by dielectric and impedance techniques. IEEE Trans Biomed Eng 2005;MBE-52:549–552.
39. http://www.patentalert.com/docs/001/z00120411.shtml.
40. Yamakoshi KI, et al. Noninvasive measurement of hematocrit by electrical admittance plethysmography technique. IEEE Trans Biomed Eng 1989;27:156–161.
41. Mishra NN, et al. Bio-impedance sensing device (BISD) for detection of human CD4+ cells. Nanotech 2004, vol. 1, Proc 2004 NSTI Nanotechnology Conf 2004, p 228–231.
42. Brouillette RT, Morrow AS, Weese-Mayer DE, Hunt CE. Comparison of respiratory inductive plethysmography and thoracic impedance for apnea monitoring. J Ped 1987;111: 377–383.
43. Valta P, et al. Evaluation of respiratory inductive plethysmography in the measurement of breathing pattern and PEEP-induced changes in lung volume. Chest 1992;102:234–238.
44. Cohen KP, et al. Design of an inductive plethysmograph for ventilation measurement. Physiol Meas 1994;15:217–229.
45. Strömberg TLT. Respiratory Inductive Plethysmography. Linköping Studies in Science and Technologies Dissertations No. 417, 1996.
46. Malmivuo J, Plonsey R. Impedance plethysmography. In: Malmivuo J, Plonsey R, editors, Bioelectromagnetism. New York: Oxford University Press; 1995. Chapt. 25.

Further Reading

Foster KR, Schwan HP. Dielectric properties of tissue and biological materials. A critical review. Crit Rev Biomed Eng 1989;17:25–102.

Kaindl F, Polzer K, Schuhfried F. (1958, 1966, 1979): Rheographie. Darmstadt, Dr: Dietrich Steinkopff Verlag; 1958 (1st ed), 1966 (2nd ed), 1979 (3rd ed).

Morucci J-P, et al. Bioelectrical impedance techniques in medicine. Crit Rev Biomed Eng 1996;24:223–681.

Pethig R. Dielectric and Electronic Properties of Biological Materials. Chichester: Wiley; 1979.

Schwan HP. Determination of biological impedances. In: Nastuk WL, editor. Physical Techniques in Biological Research. Vol. VI (ptB) New York: Academic Press; 1963; p 323–407.

Schwan HP. Biomedical engineering: A 20th century interscience. Its early history and future promise. Med Biol Eng Comput 1999;37(Suppl. 2):3–13.

Stuchly MA, Stuchly SS. Dielectric properties of biological substances—tabulated. J Microw Power 1980;15:19–26.

Webster JG, editor. Medical Instrumentation—Application and Design. 3rd ed. New York: Wiley; 1998.

Gabriel C, Gabriel S. Compilation of the Dielectric Properties of Body Tissues at RF and Microwave Frequencies, 1996. Available at http://www.brooks.af.mil/AFRL/HED/hedr/reports/dielectric/Report/Report.html.

See also Bioimpedance in cardiovascular medicine; peripheral vascular noninvasive measurements.

IMPEDANCE SPECTROSCOPY

Birgitte Freiesleben De Blasio
Joachim Wegener
University of Oslo
Oslo, Norway

INTRODUCTION

Impedance spectroscopy (IS), also referred to as electrochemical impedance spectroscopy (EIS), is a versatile approach to investigate and characterize dielectric and conducting properties of materials or composite samples (1). The technique is based on measuring the impedance (i.e., the opposition to current flow) of a system that is being excised with weak alternating current or voltage. The impedance spectrum is obtained by scanning the sample impedance over a broad range of frequencies, typically covering several decades.

In the 1920, researchers began to investigate the impedance of tissues and biological fluids, and it was early known that different tissues exhibit distinct dielectric properties, and that the impedance undergoes changes during pathological conditions or after excision (2,3).

The advantage of IS is that it makes use of weak amplitude current or voltage that ensures damage-free examination and a minimum disturbance of the tissue. In addition, it allows both stationary and dynamic electrical properties of internal interfaces to be determined, without adversely affecting the biological system. The noninvasive nature of the method combined with its high information potential makes it a valuable tool for biomedical research and many medical applications are currently under investigation and development; this will be reviewed at the end of the article.

This article starts with providing a general introduction to the theoretical background of IS and the methodology connected to impedance measurement. Then, the focus will be on applications of IS, particularly devises for *in vitro* monitoring of cultured cell systems that have attracted widespread interest due to demand for noninvasive, marker-free, and cost-effective methods.

THEORY

Impedance, $\boldsymbol{Z}$, is a complex-valued vector that describes the ability of a conducting medium to resist flow of alternating current (ac). In a typical IS experiment (Fig. 1), a sinusoidal current $\boldsymbol{I(t)}$ signal with angular frequency ω (ω = 2π*f*) is passed through the sample and the resulting steady-state voltage $\boldsymbol{U(t)}$ from the excitation is measured. According to the ac equivalent of Ohm's law, the impedance is given by the ratio of these two quantities

$$Z = \frac{U(t)}{I(t)} \tag{1}$$

The impedance measurement is conducted with use of weak excitation signals, and in this case the voltage response will be sinusoid at the same frequency ω as the applied current signal, but shifted in phase φ. Introducing complex notation, Eq. 1 translates into

$$Z = \frac{U_0}{I_0}\exp(i\varphi) = |Z|\exp(i\varphi) \tag{2}$$

with U_0 and I_0 being the amplitudes of the voltage and current, respectively, and $i = \sqrt{-1}$ being the imaginary unit. Thus, at each frequency of interest the impedance is described by two quantities: (*1*) its magnitude $|Z|$, which is the ratio of the amplitudes of $\boldsymbol{U(t)}$ and $\boldsymbol{I(t)}$; and (*2*) the phase angle φ between them. The impedance is measured over a range of frequencies between hertz and gigahertz, dependent on the type of sample and the problem to study.

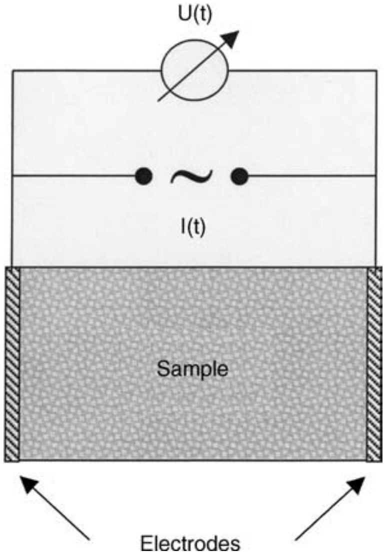


Figure 1. Schematic of a two-electrode setup to measure the frequency-dependent impedance of a sample that is sandwiched between two parallel plate electrodes.

The measured impedance can be divided into its real and imaginary components, that is, the impedance contribution arising from current in-phase with the voltage and 90° out-of-phase with the voltage

$$R = \mathrm{Re}(Z) = |Z|\cos(\varphi), \quad X = \mathrm{Im}(Z) = |Z|\sin(\varphi) \tag{3}$$

The real part is called resistance, R, and the imaginary part is termed reactance, X. The reactive impedance is caused by presence of storage elements for electrical charges (e.g., capacitors in electrical circuit).

In some cases it is convenient to use the inverse quantities, which are termed admittance $\boldsymbol{Y} = 1/\boldsymbol{Z}$, conductance $\boldsymbol{G} = \mathrm{Re}(\boldsymbol{Y})$, and susceptance $\boldsymbol{B} = \mathrm{Im}(\boldsymbol{Y})$, respectively. In the linear regime (i.e., when the measured signal is proportional to the amplitude of the excitation signal), these two representation are interchangeable and contain the same information. Thus, IS is also referred to as admittance spectroscopy.

INSTRUMENTATION

The basic devices for conducting impedance measurements consist of a sinusoid signal generator, electrodes, and a phase-sensitive amplifier to record the voltage or current. Commonly, a four-electrode configuration is used, with two current injecting electrodes and two voltage recording electrodes to eliminate the electrode–electrolyte interface impedance. As discussed below, some applications of IS make use of two-electrode arrangements in which the same electrodes are used to inject current and measure the voltage.

Since the impedance is measured by a steady-state voltage during current injection, some time is needed when changing the frequency before a new measurement can be performed. Therefore, it is very time consuming if each frequency has to be applied sequentially. Instead, it is common to use swept sine wave generators, or spectrum analyzers with transfer function capabilities and a white noise source. The white noise signal consists of the superposition of sine waves for each generated frequency, and the system is exposed to all frequencies at the same time. Fourier analysis is then used to extract the real and imaginary parts of the impedance.

The electrodes used for impedance experiments are made from biocompatible materials, such as noble metals, which in general is found not to have deleterious effect on biological tissue function. Electrode design is an important and complicated issue, which depends on several factors, including the spatial resolution required, the tissue depth, and so on. It falls beyond the scope of this article to go further into details. The interested readers are referred to the book by Grimnes and Martinsen (4) for a general discussion.

Common error sources in the measurements include impedance drift (e.g., caused by adsorption of particles on the electrodes or temperature variations). Ideally, the system being measured should be in steady-state throughout the time required to perform the measurement, but in practice this can be difficult to achieve. Another typical error source is caused by pick-up of electrical noise from the electronic equipment, and special attention must be paid to reduce the size of electric parasitics arising, for example, from cables and switches.

DATA PRESENTATION AND ANALYSIS

The most common way to analyze the experimental data is by fitting an equivalent circuit model to the impedance spectrum. The model is made by a collection of electrical elements (resistors, capacitors) that represents the electrical composition in the system under study.

As a first step, it is useful to present the measured data by plotting log $|Z|$ and φ versus log f in a so-called Bode-diagram (Fig. 2a), and by plotting Im$|Z|$ versus Re$|Z|$ named a Nyquist diagram or an *impedance locus* (Fig. 2b). The examples provided in Fig. 2 are made for an electrical circuit (insert Fig. 2b). While the first way of presenting the data shows the frequency-dependence explicitly, the phase angle φ is displayed in the latter.

The impedance spectrum gives many insights to the electrical properties of the system, and with experience it is possible to make a qualified guess of a proper model based on the features in the diagrams (cf. Fig. 4). Similar to other spectroscopic approaches like infrared (IR) or ultraviolet (UV)/visible (vis), the individual components tend to show up in certain parts of the impedance spectrum. Thus, variations in the values of individual components alter the spectrum in confined frequency windows (Fig. 3).

For a given model, the total impedance (transfer function) is calculated from the individual components with use of Ohm's and Kirchhoff's laws. The best estimates for the parameters, that is, the unknown values of the resistors and capacitors in the circuit, are then computed with use of least-square algorithms. If the frequency response of the chosen model fits the data well, the parameter values are used to characterize the electrical properties of the system.

In order to fit accurately the equivalent circuit model impedance to the impedance of biomaterials, it is often necessary to include nonideal circuit elements, that is, elements with frequency dependent properties. Such elements are not physically realizable with standard technical elements. Table 1 provides a list of common circuit elements that are used to describe biomaterials with respect to their impedance and phase shift. The constant phase element (CPE) portrays a nonideal capacitor, and was originally introduced to describe the interface impedance of noble metal electrodes immersed in electrolytic solutions (5). The physical basis for the CPE in living tissue (and at electrode interfaces) is not clearly understood, and it is best treated as an empirical element. Another example is the Warburg impedance σ that accounts for the diffusion limitation of electrochemical reactions (4).

It is important to place a word of caution concerning the equivalent circuit modeling approach. Different equivalent circuit models (deviating with respect to components or in the network structure) may produce equally good fits to the experimental data, although their interpretations are very different. It may be tempting to increase the number of elements in a model to get a better agreement between experiment and model. However, it may then occur that the model becomes redundant because the components cannot be quantified independently. Thus, an overly complex

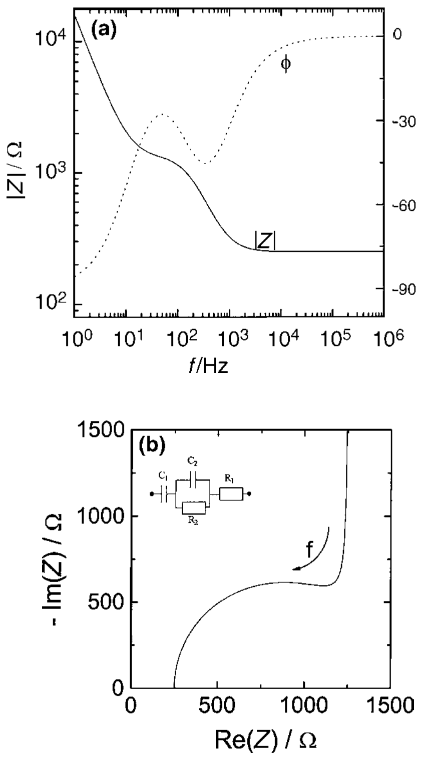


Figure 2. Different representations of impedance spectra. (a and b) visualize the frequency-dependent complex impedance of the electrical circuit shown in (b) with the network components: $R_1 = 250\ \Omega$; $R_2 = 1000\ \Omega$; $C_1 = 10\ \mu\text{F}$; $C_2 = 1\ \mu\text{F}$. (a) Bode-diagram presenting frequency-dependent impedance magnitude $|Z|$ together with the phase shift φ of the sample under investigation. (b) Wessel diagram locus of the same electrical network. The imaginary component of the impedance (reactance) is plotted against the real component (resistance). The arrow indicates the direction along which the frequency increases.

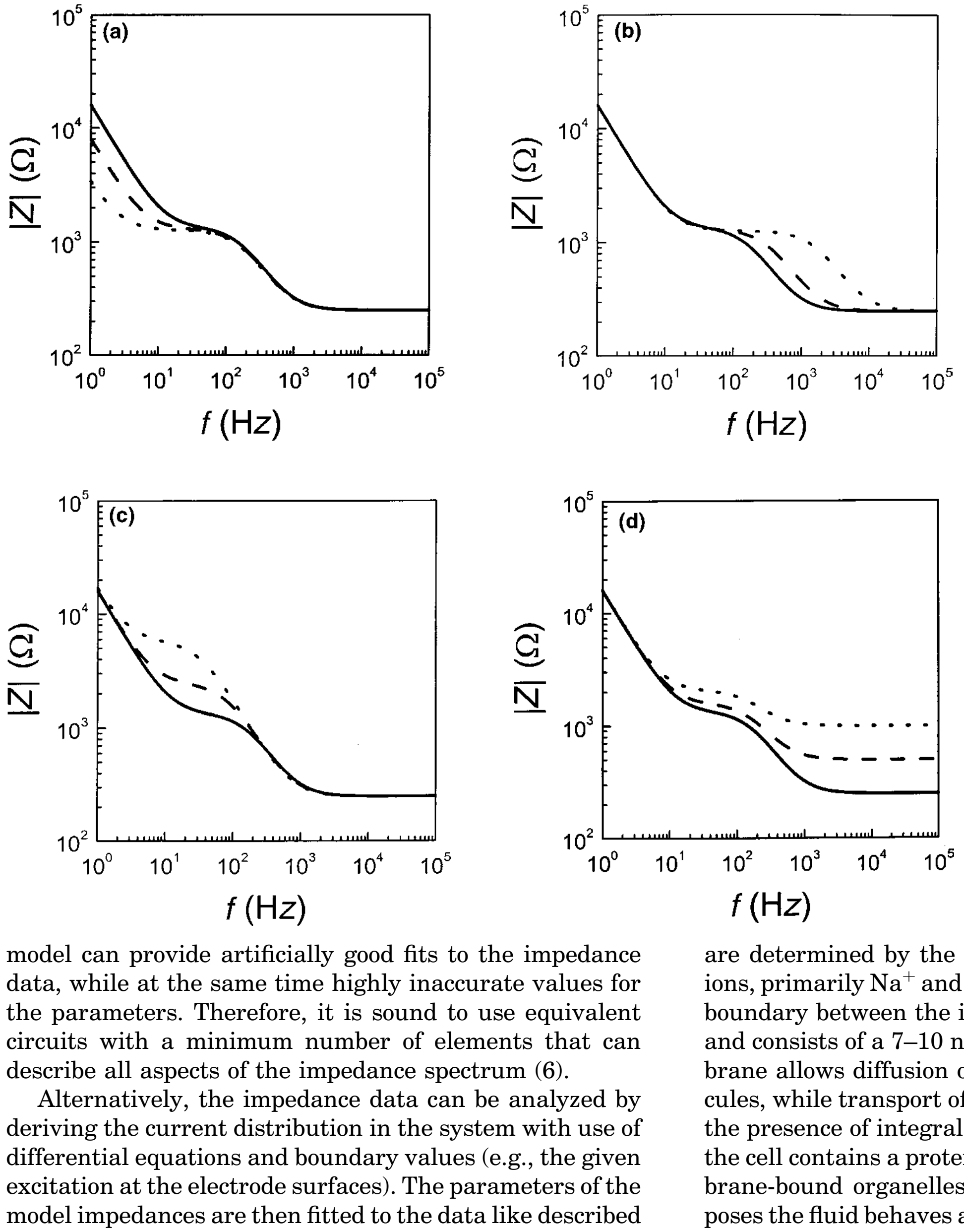


Figure 3. Calculated spectra of the impedance magnitude $|Z|$ for the electrical circuit shown in the insert of Fig. 2b when the network parameters have been varied individually. The solid line in each figure corresponds to the network parameters: $R_1 = 250$ Ω; $R_2 = 1000$ Ω; $C_1 = 10$ μF; $C_2 = 1$ μF. (a) Variation of C_1: 10 μF (solid), 20 μF (dashed), 50 μF (dotted). (b) Variation of C_2: 1 μF (solid), 0.5 μF (dashed), 0.1 μF (dotted). (c) Variation of R_1: 1000 Ω (solid), 2000 Ω (dashed), 5000 Ω (dotted). (d) Variation of R_2: 250 Ω (solid), 500 Ω (dashed), 1000 Ω (dotted).

model can provide artificially good fits to the impedance data, while at the same time highly inaccurate values for the parameters. Therefore, it is sound to use equivalent circuits with a minimum number of elements that can describe all aspects of the impedance spectrum (6).

Alternatively, the impedance data can be analyzed by deriving the current distribution in the system with use of differential equations and boundary values (e.g., the given excitation at the electrode surfaces). The parameters of the model impedances are then fitted to the data like described above. An example of this approach is be presented below, where it is used to analyze the IS of a cell-covered gold film electrode.

IMPEDANCE ANALYSIS OF TISSUE AND SUSPENDED CELLS

The early and pioneering work on bioimpedance is associated with the names of Phillipson, et al. (5). In these studies blood samples or pieces of tissue were examined in an experimental setup as shown in Fig. 1, and the dielectric properties of the biological system were investigated over a broad range of frequencies from hertz to gigahertz.

To understand the origin of bioimpedance, it is necessary to look at the composition of living material. Any tissue is composed of cells that are surrounded by an extracellular fluid. The extracellular medium contains proteins and polysaccharides that are suspended in an ionic solution and the electrical properties of this fluid are determined by the mobility and concentration of the ions, primarily Na^+ and Cl^-. The cell membrane marks the boundary between the interior and exterior environment, and consists of a 7–10 nm phospholipid bilayer. The membrane allows diffusion of water and small nonpolar molecules, while transport of ions and polar molecules requires the presence of integral transport proteins. On the inside, the cell contains a protein-rich fluid with specialized membrane-bound organelles, like the nucleus. For most purposes the fluid behaves as a pure ionic conductor. Thus, the cell membrane is basically a thin dielectric sandwiched between two conducting media and in a first approximation its impedance characteristics are mainly capacitive.

The simplest possible explanatory model for biological tissue (Fig. 4a-1) therefore consists of two membrane capa-

Table 1. Individual Impedance Contributions of Ideal and Empirical Equivalent Circuit Elements

Component of Equivalent Circuit	Parameter	Impedance, Z	Phase Shift, φ
Resistor	R	R	0
Capacitor	C	$(i\omega C)^{-1}$	$-\pi/2$
Coil	L	$i\omega L$	$+\pi/2$
Constant phase element (CPE)	$\alpha (0 \leq \alpha \leq 1)$	$1/(iC\omega)\alpha$ ·	$-\alpha\pi/2$
Warburg impedance, σ	σ	$\sigma(1 - i)\omega^{-0.5}$	$-\pi/4$

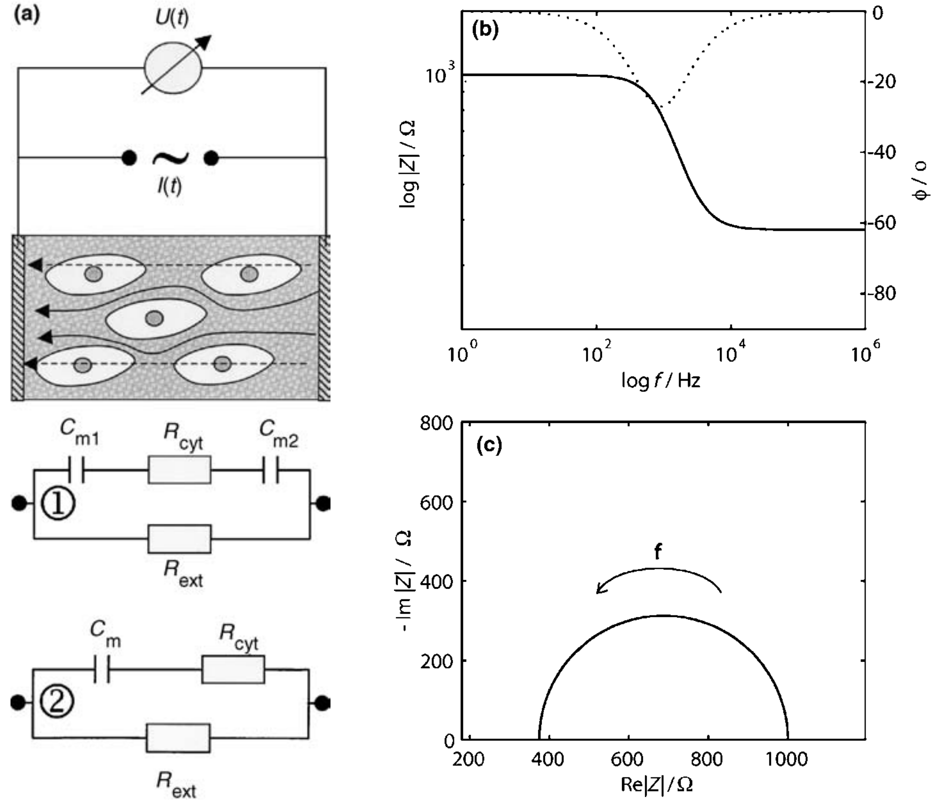


Figure 4. (a) Schematics of the impedance measurement on living tissue. The arrows indicate the pathway of current flow for low frequencies (solid line) and high frequencies (dashed). Only at high frequencies the current flows through the cells. The electrical structure of tissue can be directly translated into equivalent circuit **1**, which can be simplified to equivalent circuit **2**. (b) Bode-diagram for network 2 in Fig. 4a, using $R_{\text{ext}} = 1000\ \Omega$; $R_{\text{cyt}} = 600\ \Omega$; $C_{\text{m}} = 100$ nF. (c) Impedance locus generated with the same values.

citors C_{m1} and C_{m2} in series with a resistor of the cytosolic medium R_{cyt}, which in turn acts in parallel with the resistance of the conducting extracellular medium R_{ext}. Since the two series capacitances of the membrane cannot be determined independently, they are combined to an overall capacitance C_{m}. The equivalent circuit model of this simple scenario (Fig. 4a-2) gives rise to a characteristic impedance spectrum, as shown with the Bode-diagram (Fig. 4b) and the Nyquist diagram (Fig. 4c). Impedance data for biological tissue is also often modeled by the so-called Cole–Cole equation (7)

$$Z = R_\infty + \frac{\Delta R}{1 + \Delta R (i\omega C)^\alpha} \quad \Delta R = R_0 - R_\infty \tag{4}$$

This simple empirical model is identical to the circuit of Fig. 4, except that the capacitor is replaced by a CPE element. The impedance spectrum is characterized by four parameters $(\Delta R, R_\infty, \alpha, \tau)$, where R_0, R_∞ is the low- and high frequency intercepts on the x axis in the Nyquist plot (cf. Fig. 4d), τ is the time constant $\tau = \Delta R \cdot C$, and α is the CPE parameter. The impedance spectrum will be similar to Fig. 4b, c, but when $\alpha \neq 1$, the semicircle in the Nyquist diagram is centered below the real axis, and the arc will appear flattened. For macroscopically heterogeneous biological tissue, the transfer function is written as a sum of Cole–Cole equations.

The features of the impedance spectrum Fig. 4b can be intuitively understood: at low frequencies the capacitor prevents current from flowing through R_{cyt} and the measured impedance arises from R_{ext}. At high frequencies, with the capacitor having a very low impedance, the current is free to flow through both R_{cyt}, R_{ext}. Thus, there is a transition from constant-level impedance at low frequencies to another constant level. This phenomenon is termed dispersion, and will be discussed in the following.

A homogenous conducting material is characterized by a bulk property named the resistivity ρ' having the dimensions of ohm centimeters (Ω·cm). Based on this intrinsic parameter, the resistance may be defined by

$$R = \frac{\rho' L}{A} \tag{5}$$

where A is the cross-sectional area and L is the length of the material. Thus, by knowing the resistivity of the material and the dimensions of the system being studied, it is possible to estimate the resistance. Similarly, a homogeneous dielectric material is characterized by an intrinsic property called the relative permittivity ε', and the capacitance is defined by

$$C = \frac{\varepsilon' \varepsilon_0 A}{d} \tag{6}$$

where ε_0 is the permittivity of free space with dimension F/m, and A, d are the dimensions of the system as above. For most biological membranes, the area-specific capacitance is found to be quite similar, with a value of $\sim 1\ \mu\text{F} \cdot \text{cm}^{-2}$ (8).

For historical reasons the notation of conductivity σ' with dimensions S·m^{-1} and conductance $(G = \sigma' A/d)$ has been preferred over resistance R and resistivity ρ, but the information content is the same, it is just expressed in a different way.

It is possible to recombine ε' and σ' by defining a complex permittivity $\varepsilon = \varepsilon' + \varepsilon''$, with $\text{Re}(\varepsilon) = \varepsilon'$ and $\text{Im}(\varepsilon) = \varepsilon''$. The imaginary part accounts for nonideal capacitive behavior, for example, current within the dielectric due to bound charges giving rise to a viscous energy loss (dielectric loss). Therefore, ε″ is proportional to σ′, when adjusted for the conductivity that is due to migration σ_0 (9)

$$\varepsilon'' = \frac{\sigma' - \sigma_0}{2\pi f \varepsilon_0} \tag{7}$$

When a piece of biological material is placed between two electrodes, it is possible to measure the capacitance of the system and thereby to estimated the tissue permittivity ε′. In general, ε′ quantifies the ratio of the capacitance when a dielectric substance is placed between the electrodes, relative to the situation with vacuum in between. The increase of capacitance upon insertion of a dielectric material is due to polarization in the system in response to the electric field. For direct current (dc) or low frequency situations ε′ is called the dielectric constant. When the frequency is increased, ε′ often shows strong frequency dependence with a sigmoid character in a log–log plot of ε′ versus frequency. This step-like decrease of the permittivity is referred to as a dielectric dispersion. The frequency f_c at which the transition is half-complete is called the characteristic frequency, and is often expressed as time constant τ with

$$\tau = \frac{1}{f_c} \tag{8}$$

Going back to Fig. 4c, the characteristic frequency is found directly as the point when the phase angle is at maximum.

The origin of dielectric dispersion in a homogeneous material is due to a phenomenon termed orientation polarization. Dipolar species within the material are free to move and orient themselves along the direction of the field, and therefore they contribute to the total polarization. However, when the frequency becomes too high, the dipoles can no longer follow the oscillation of the field, and their contribution vanishes. This relaxation causes the permittivity ε′ to decrease.

For heterogeneous samples like tissue additional relaxation phenomena occur, leading to more complex frequency dependence. In 1957, Schwan (10) defined three prominent dispersion regions of relevance for bioimpedance studies called α, β, and γ, which is shown in Fig. 5. The dispersions are generally found in all tissue, although the time constant and the change in permittivity Δε′ between the different regions may differ (9).

Briefly stated, the α-dispersion originates from the cloud of counterions that are attracted by surface charges of the cell membrane. The counterions can be moved by an external electric field, thereby generating a dipole moment and relaxation. The β-dispersion, which is also called Maxwell–Wagner dispersion, is found in a window between kilohertz and megahertz (kHz and MHz). It arises due to the accumulation of charges at the interface between hydrophobic cell membranes and electrolytic solutions. Since the aqueous phase is a good conductor, whereas the membrane is not, mobile charges accumulate and charge up the membrane capacitor, thus, contributing to polarization. When the frequency gets too high, the charging is not complete, causing a loss of polarization. Finally, the γ-dispersion is due to the orientation polarization of small molecules, predominantly water molecules.

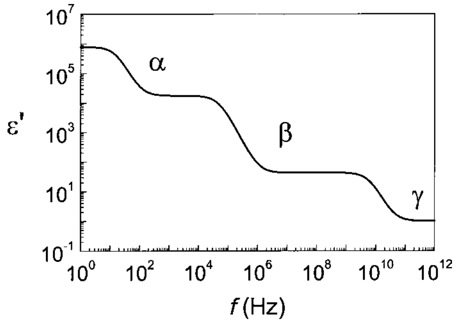


Figure 5. Frequency-dependent permittivity ε′ of tissue. The permittivity spectrum ε′(*f*) is characterized by three major dispersions: α-, β-, and γ-dispersion.

Most IS measurements are performed at intermediate frequencies in the regime of the β-dispersion. In this frequency window, the passive electrical properties of tissue are well described with the simple circuit shown in Fig. 4 or by the Cole–Cole equation. The measurements can be used to extract information about extra- and intercellular resistance, and membrane capacitance. For example, it has been shown that cells in liver tissue swell, when the blood supply ceases off (ischemia) and that the swelling of the cells can be monitored as an increase in the resistance of the extracellular space R_{ext} (11). Cell swelling compresses the extracellular matrix around the cells, and thereby narrows the ion pathway in this region. Based on experiments like these, there is a good perspective and prognosis that IS may serve as a routine monitoring tool for tissue vitality even during the surgery.

APPLICATION: MONITORING OF ADHERENT CELLS *IN VITRO*

The attachment and motility of anchorage dependent cell cultures is conveniently studied using a microelectrode setup. In this technique, cells are grown directly on a surface containing two planar metal electrodes, one microelectrode and one much larger counter electrode. The cells are cultured in normal tissue culture medium that serves as the electrolyte.

When current flows between the two electrodes, the current density, and the measured voltage drop, will be much higher at the small electrode. Therefore the impedance measurement will be dominated by the electrode polarization of the small electrode Z_{el}. Instead, no significant polarization takes place at the larger counter electrode and its contribution to the measured impedance may be ignored. The electrode polarization impedance Z_{el} acts physically in series with the resistance of the solution R_{sol}. Since the current density is high in a zone (the constrictional zone) proximal to the microelectrode, the electrolytic resistance will be dominated by the constriction resistance R_c in this region (Fig. 6). The total measured impedance may therefore be approximated by $\mathbf{Z} \sim \mathbf{Z}_{\text{el}} + R_c$ (4). If necessary, R_c may be determined from high frequency measurements where the electrode resistance is

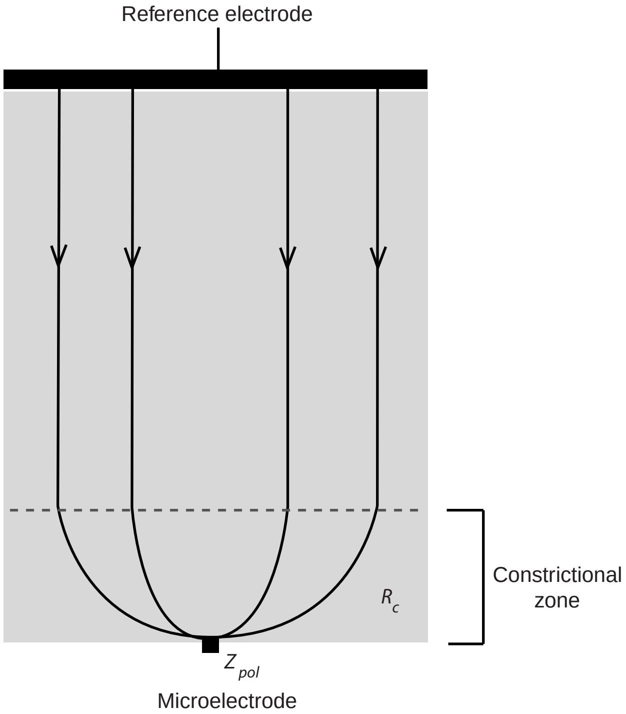


Figure 6. Schematic of two-electrode configuration.

infinitely small, that is, $\mathrm{Re}(Z_{\mathrm{el}}) \to 0$, and subtracted from the measured impedance to determine the impedance of the electrode–electrolyte interface.

When cells adhere on the small electrode, they constrain the current flow from the interface, increasing the measured impedance. The changes mainly reflects the capacitive nature of nonconducting lipid-bilayer membrane surrounding the cells. The cell membranes cause the current field to bend around the cells, much like if they were microscopic insulating particles. It is possible to follow both cell surface coverage and cell movements on the electrode, and morphological changes caused by physiological/pathological conditions and events may be detected. The technique may also be used to estimate cell membrane capacitances, and barrier resistance in epithelial cell sheets. In addition, the method is highly susceptible to vertical displacements of the cell body on the electrode with sensitivity in the nanometer range.

INSTRUMENTATION

The technique was introduced by Giaever and Keese in 1984 and referred to as Electrical Cell-Substrate Impedance Sensing (ECIS) (12,13). The ECIS electrode array consists of a microdisk electrode ($\sim 5 \times 10^{-4}\,\mathrm{cm}^2$) and a reference electrode ($\sim 0.15\,\mathrm{cm}^2$); depending on the cell type to be studied, the recording disk electrode may contain a population of 20–200 cells. The electrodes are made from depositing gold film on a polycarbonate substrate over which an insulating layer of photoresist is deposited and delineated. A 1 V amplitude signal at fixed frequency (0.1–100 kHz) is applied to the electrodes through a large resistor to create a controlled current of $1\,\mu A$, and the corresponding voltage across the cell-covered electrodes is measured by a lock-in amplifier, which allows amplification of the relatively small signals. The amplifier is interfaced to a PC for data storage. The impedance is calculated from the measured voltage displayed in real time on the computer screen (Fig. 7). During the measurements the sample is placed in an incubator at physiological conditions.

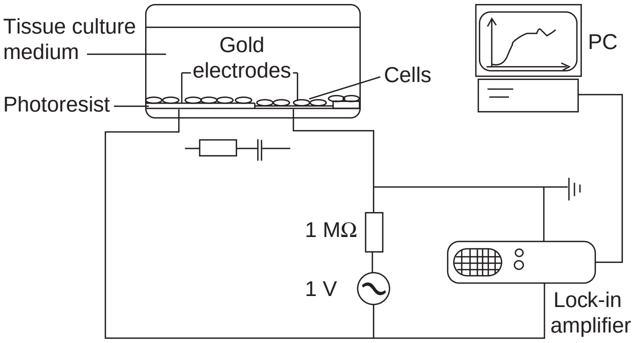


Figure 7. The ECIS measurement setup.

The ECIS system is now commercially available, and the electrode slides allow multiple experiments to be performed at the same time (14). Some modifications to the technique have been described, such as a two-chamber sample well, which permit simultaneous monitoring on a set of empty electrodes being exposed to the same solution (15), platinized single-cell electrodes (15), and inclusion of a voltage divider technique to monitor the impedance across a range of frequencies (16). More recently, impedance studies have been performed using other types of electrode design. One approach has been to insert a perforated silicon-membrane between two platinium electrodes, there by allowing for two separate electrolytic solutions to exist on either side of the membrane (17). The results obtained with these techniques are generally identical to those obtained by the ECIS system.

MODEL OF ELECTRODE–CELL INTERFACE

To interpret ECIS-based impedance data, a model of the ECIS electrode–cell interface has been developed that allows determination of (*1*) the distance between the ventral cell surface and the substratum, (*2*) the barrier resistance, and (*3*) the cell membrane capacitance of confluent cell layers (18). The model treats the cells as disk shaped objects with a radius r_c that are separated an average distance h from the substrate (Fig. 8). When cells cover the electrode, the main part of the current will flow through the thin layer of medium between the cell and the electrode, and leave the cell sheet in the narrow spacing between cells. However, the cell membrane, which is modeled as a capacitor (an insulating layer separating the conducting fluids of the solution and the cytosol) allows a parallel current flow to pass through the cells. The minor resistive component of the membrane impedance due to the presence of ionic channels is ignored in the calculations. By assuming that the electrode properties are not affected by the presence of cells, a boundary-value model of the current flow across the cell layer may be used to derive a relation

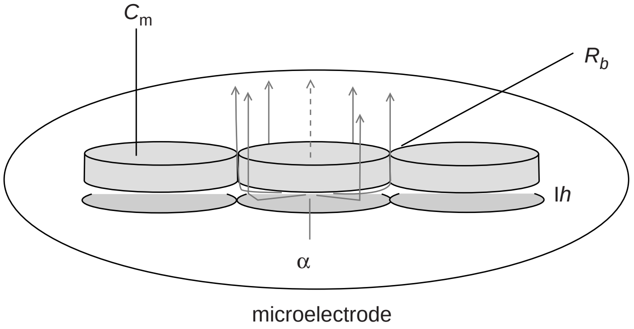


Figure 8. Model of current flow paths. The impedance changes associated with the presence of cells, arise in three different regions: from current flow under the cells quantified by α, from current flow in the narrow intercellular spacings causing the barrier resistance R_b. In parallel, some current will pass through the cell membranes giving rise to capacitive reactance C_m.

between the specific impedance of a cell-covered electrode Z_{cell} and the empty electrode Z_{el}

$$\frac{1}{Z_{\mathrm{cell}}} = \frac{1}{Z_{\mathrm{el}}}\left(\frac{Z_{\mathrm{el}}}{Z_{\mathrm{el}}+Z_{\mathrm{m}}} + \frac{\frac{Z_{\mathrm{m}}}{Z_{\mathrm{el}}+Z_{\mathrm{m}}}}{\frac{\gamma r_{\mathrm{c}}}{2}\frac{I_0(\gamma r)}{I_1(\gamma r)} + R_{\mathrm{b}}\left(\frac{1}{Z_{\mathrm{el}}}+\frac{1}{Z_{\mathrm{m}}}\right)}\right)$$

$$\gamma = \sqrt{\frac{\rho}{h}\left(\frac{1}{Z_{\mathrm{el}}}+\frac{1}{Z_{\mathrm{m}}}\right)} \qquad (9)$$

where I_0, I_1 are the modified Bessel functions of the first kind of order zero and one, R_b and ρ are the specific barrier resistance and resistivity of the solution, and $Z_m = -2i/(\omega C_m)$ is the specific membrane impedance of the cells. A parameter $\alpha = r_c(\rho/h)^{0.5}$ is introduced as an assessment of the constraint of current flow under the cells. The impedance spectrum of an empty electrode and a cell-covered electrode is used to fit the three adjustable parameters (R_b, α, C_m).

The model outlined above has been further refined to describe polar epithelial cell sheets, treating separately the capacitance of the apical, basal, and lateral membranes (19). Some applications of the model will be discussed in the following sections.

MONITORING ATTACHMENT AND SPREADING

As a cell comes into contact with a solid surface, it forms focal contacts, primarily mediated by transmembrane proteins that anchor structural filaments in the cell interior to proteins on the substrate. During this process, the rounded cell spreads out and flattens on the surface, greatly increasing its surface area in contact with the electrode. The cell will also establish contacts with neighboring cells through particular cell–cell junctions, such as tight junctions, where strands of transmembrane proteins sew neighboring cells together, and gap junctions formed by clusters of intercellular channels, connecting the cytosol of adjacent cells.

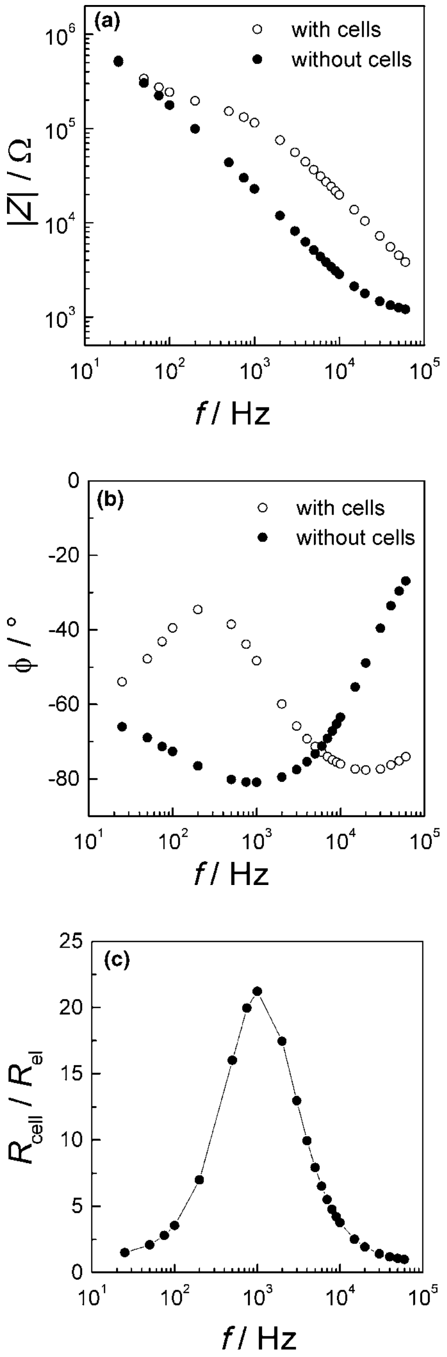


Figure 9. (a, b) Bode-diagrams of an ECIS electrode with confluent MDCK cells and an empty electrode. (c) Plot showing the division of the measured resistance of an ECIS electrode with confluent MDCK cells with the corresponding values of the empty electrode plotted versus log *f*.

The attachment process is normally studied using single-frequency measurements. Figure 9a and b show Bode plots for ECIS data of an empty electrode and an electrode with a confluent layer of epithelial MDCK cells. It is seen that the presence of cells primarily affects the impedance spectrum for intermediate frequencies between 1 and 100 kHz (Fig. 9a). At the highest frequencies, the two plots approach a horizontal plateau that represents the ohmic solution resistance between the working and the counter electrode. Within the relevant frequency window, the phase-shift plot for the data of the cell-covered electrode displays two extrema. At frequencies ~ 200 Hz, the phase shift φ is closest to zero, indicating that the contribution of the cells on the measured impedance is mainly resistive. At higher frequencies, the effect of the cell layer becomes more capacitive, and φ starts approaching −90°. The impedance spectrum of the empty electrode displays a single dispersion related to double-layer capacitance at the electrode interface.

The ideal measurement frequencies, where the presence of cells is most pronounced, are determined by dividing the impedance spectrum of a cell-covered electrode with the spectrum of a naked electrode. The same can be done for the resistance or capacitance spectrum, respectively. The most sensitive frequency for resistance measurements is typically found between 1 and 4 kHz (Fig. 9c), where the ratio $R_{\mathrm{cell}}(f)\ /\ R_{\mathrm{el}}(f)$ is at maximum. The capacitive contribution peaks at much larger frequencies, typically on the order of 40 kHz, so that capacitance measurements are often performed at this higher frequency.

During the initial hours following the initial electrode–cell contact, the monitored impedance undergoes a characteristic steep increase. Once the spreading is complete, the impedance continues to fluctuate, reflecting the continuous morphological activities of living cells, for example, movements of cells on the electrode, either by local protrusions or directed movements of the entire cell body, or cell divisions (Fig. 10). The signal characteristics of the impedance during the spreading phase are generally found to be distinct for different cell cultures, both in terms of the duration of the initial gradient and its relative size in comparison to the impedance recorded from a the naked electrode (20). Also, characteristic impedance curves can be obtained by coating the electrode with different proteins (e.g., fibronectin, vitronectin) (21).

Simultaneous optical monitoring of a transparent ECIS electrode has allowed systematic comparison of cell confluence and measured impedance (22). Analysis of data from subconfluent MDCK epithelial cultures revealed a strong linear association between the two variables with cross-correlation coefficients > 0.9; the correlation was found to be equally strong in early and late cultures. This result indicates that ~ 80% of the variance in the measured resistance (4 kHz) can be attributed to changes in cell coverage area (Fig. 11). Moreover, it was possible to link resistance variations to single-cell behavior during cell attachment, including cell-division (temporary impedance plateau) and membrane ruffling (impedance increase). The measured cell confluence was compared to the theoretical model (Eq. 9), neglecting the barrier resistance (i.e., $R_{\mathrm{b}} = 0$), and the calculated values were found to agree well with the data (Fig. 12). Studies like these might pave the way for standardized use of ECIS to quantify attachment and spreading of cell cultures.

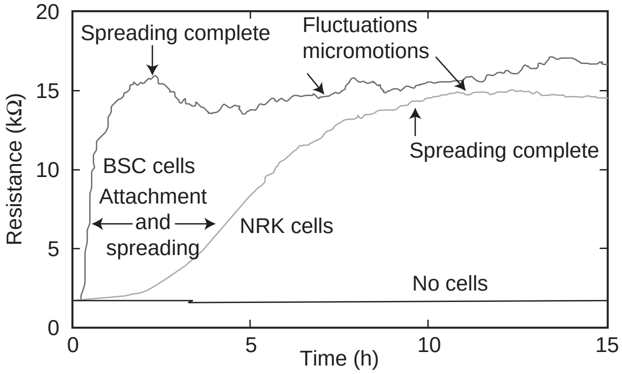


Figure 10. Attachment assay of BSC and NRK fibroblastic cells followed for an interval of 15 h. The graph shows the measured resistance (4 kHz) as function of time; the spreading phase is indicated with arrows.

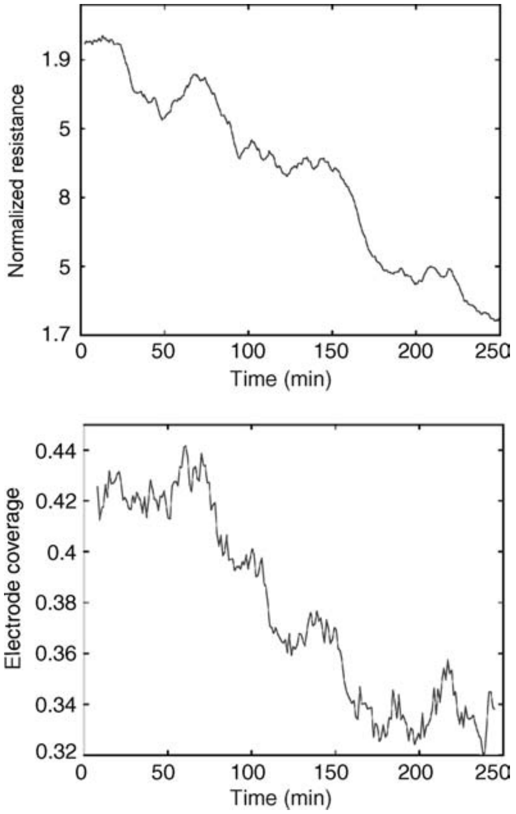


Figure 11. Correlation between resistance and cell coverage. The normalized resistance (4 kHz) versus time (upper panel), and the electrode coverage versus time (lower panel) during the same time interval. The measurement was started 32 h after the cells had been seeded out; the cross-correlation factor was $r = 0.94$.

IMPEDANCE SPECTROSCOPY AS A TRANDUCER IN CELL-BASED DRUG SCREENING

Another application of impedance spectroscopy with strong physiological and medical relevance is its use as transducer in ECIS-like experiments for cell-based drug screening assays. Here, the impedance readout can be used to monitor the response of cells upon exposure to a certain drug or a drug mixture. In these bioelectric hybrid assays the cells serve as the sensory elements and they determine the specificity of the screening assay while the electrodes are used as transducer to make the cell behavior observable. In the following example, endothelial cells isolated from bovine aorta (BAEC = bovine aortic endothelial cells) were grown to confluence on gold-film electrodes since they

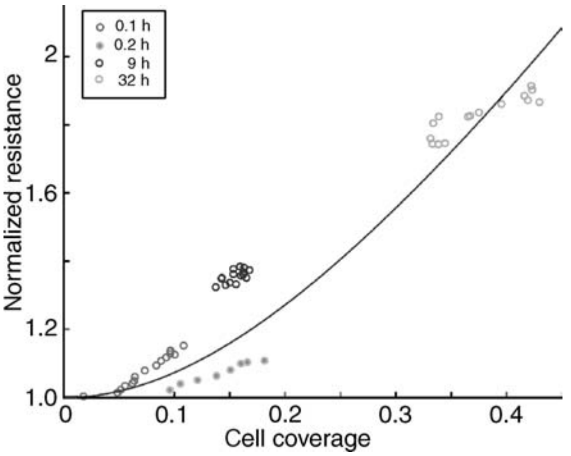


Figure 12. Theoretical prediction of cell coverage. Theoretical curve of normalized resistance plotted as function of cell coverage on the electrode. Normalized resistance and corresponding cell density are shown for four different registrations with circles. Time points indicate when the recordings were initiated with respect to the start of the culture; each circle corresponds to average values for 15 min time intervals.

express cell-surface receptors (β-adrenoceptors) that are specific for adrenalin and derivatives (23,24). These β-adrenoceptors belong to the huge family of G-protein coupled receptors (GPCR) that are of great pharmaceutical relevance and impact. By measuring the electrical impedance of the cell-covered electrode, the stimulation of the cells by the synthetic adrenaline analogue isoprenaline (ISO) can be followed noninvasively in real time without any need to apply costly reagents or to sacrifice the culture (25). Experimentally, the most sensitive frequency for time-resolved impedance measurements is first determined from a complete impedance spectrum along an extended frequency range as depicted in Fig. 13. The figure compares the impedance spectrum of a circular gold-film electrode ($d = 2$ mm) with and without a confluent monolayer of BAECs. The contribution of the cell layer to the total impedance of the system is most pronounced at frequencies close to 10 kHz, which is, thus, the most sensitive sampling frequency for this particular system. It is noteworthy that the most sensitive frequency may change with the electrode size and the individual electrical properties of the cells under study.

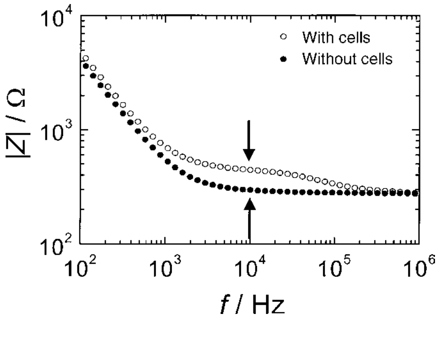


Figure 13. Frequency-dependent impedance magnitude for a planar gold-film electrode ($d = 2$ mm) with and without a confluent monolayer of BAEC directly growing on the electrode surface. The difference in impedance magnitude is maximum at a frequency of 10 kHz.

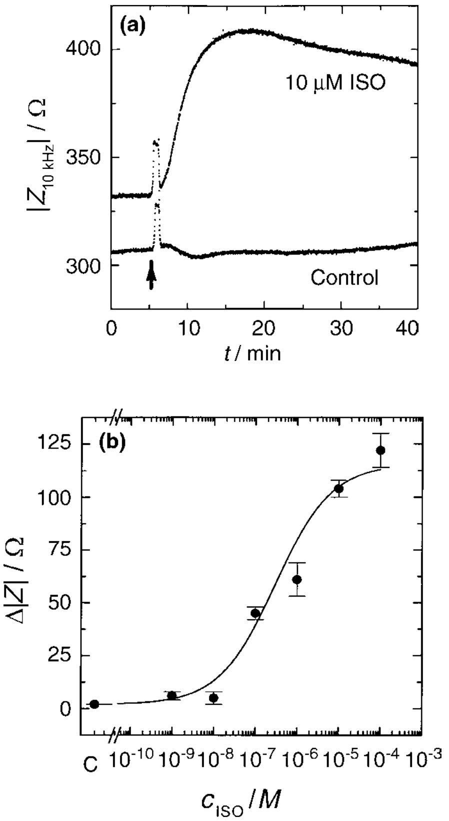


Figure 14. (a) Time course of the impedance magnitude at a sampling frequency of 10 kHz when a confluent monolayer of BAECs is exposed to 10 μ*M* ISO or a corresponding vehicle control. (b) Dose-response relationship between the increase of impedance magnitude ΔZ and the concentration of isoproterenol applied. Quantitative analysis reveals an EC_{50} of 0.3 μ*M* similar to the binding constant of ISO to β-adrenoceptors.

Figure 14a traces the time course of the impedance magnitude at a frequency of 10 kHz when confluent BAEC monolayers were either challenged with 10 μ*M* ISO or a vehicle control solution at the time indicated by the arrow. The exchange of fluids produces a transient rise of the impedance by 10–20 Ω that is not caused by any cellular response, but mirrors the reduced fluid height within the measuring chamber. As expected, no response of the cells is seen in the control experiment. The cell population exposed to 10 μ*M* of ISO shows a significant increase in electrical impedance that goes through a maximum 10 min after ISO application, and then slowly declines. The reason for the increase in impedance as observed after ISO stimulation is similar to what has been described for three-dimensional (3D) tissues above. The adrenaline derivative induces a relaxation of the cytoskeleton that in turn makes the cells flat out a bit more. As a consequence the extracellular space between adjacent cells narrows and increases the impedance of the cell layer. Note that the time resolution in these measurements is ~ 1 s so that even much faster cell responses than the one studied here can be monitored in real time. Moreover, no labeled probe had to be applied and the sensing voltages used for the measurement ($U_0 = 10$ mV) are clearly noninvasive.

From varying the ISO concentration, a dose-response relationship (Fig. 14b) can be established which is similar to those derived from binding studies using radiolabeled ligands. Fitting a dose-response transfer function to the recorded data returns the concentration of half-maximum efficiency EC_{50} as (0.3 ± 0.1) μM, which is in close agreement to the binding constant of ISO to β-adrenoceptors on the BAEC surface as determined from binding assays with radiolabeled analogs (23).

These kind of electrochemical impedance measurements are also used to screen for potent inhibitors of cell-surface receptors. Staying with the example discussed in the preceding paragraph, the blocking effect of Alprenolol (ALP), a competitive inhibitor of β-adrenoceptors (β-blocker), is demonstrated. Preincubation of BAEC with ALP blocks the stimulating activity of ISO, as shown in Fig. 15. The figure compares the time course of the impedance magnitude at a frequency of 10 kHz when BAEC monolayers were stimulated with 1 μM ISO either in absence of the β-blocker (a) or after preincubation (b). When the cell layers were incubated with 10 μM ALP prior to the addition of 1 μM ISO, the cells do not show any ISO response indicating that a 10-fold increase of ALP was sufficient to block activation of the receptors. To prove that the cells were not compromised in any way during these experiments the same signal transduction cascade was triggered via a receptor-independent way at the end of each experiment. This can be easily done by application of FOR, a membrane permeable drug that intracellularly activates the same enzyme that is triggered by ISO binding to the receptor. Forskolin stimulation of those cells that had been blocked with ALP earlier in the experiment induces a strong increase of electrical impedance indicating that the intracellular transduction pathways are functional (Fig. 15b).

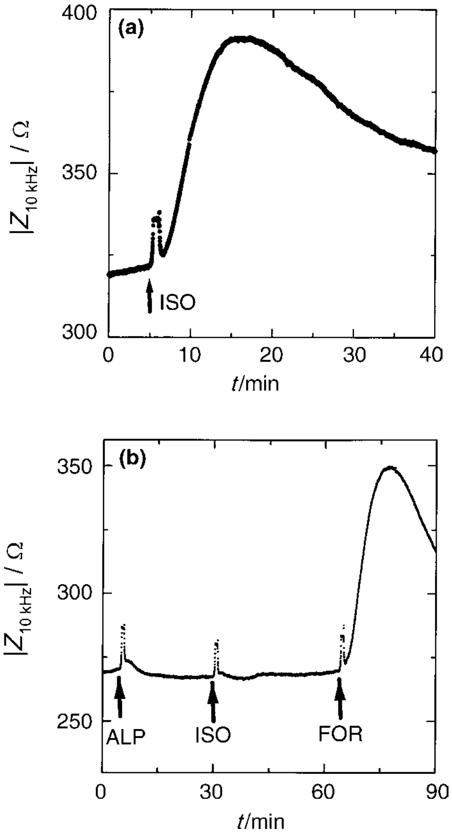


Figure 15. (a) Time course of the impedance magnitude at a sampling frequency of 10 kHz, when confluent BAEC are exposed to 1 μM ISO. (b) Time course of the impedance magnitude of a confluent monolayer of BAECs upon sequential exposure to 10 μM of the β-blocker ALP and 1 μM ISO 20 min later. The β-adrenergic impedance increase is omitted by the β-blocker. Intactness of the signal transduction cascade is verified by addition of forskolin (FOR), a receptor independent activator of this signal transduction pathway.

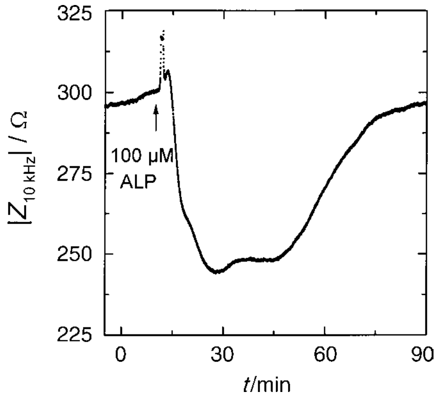


Figure 16. Time course of the impedance magnitude at a sampling frequency of 10 kHz when a confluent BAEC monolayer is exposed to an over dose of the β-blocker alprenolol (100 μM ALP). Addition of ALP is indicated by the arrow.

Besides screening for the activity of drugs in cell-based assays, these kind of measurements are also used to check for unspecific side effects of the compounds of interest on cell physiology. Dosage of ALP and many of its fellow β-blockers has to be adjusted with great care since these lipophilic compounds are known to integrate nonspecifically into the plasma membrane. As shown in Fig. 15b, application of 10 μM ALP does not show any measurable side effects. Using ALP in concentrations of 100 μM induces a transient but very pronounced reduction of the electrical impedance (Fig. 16). This decrease in impedance may be the result of the interaction of ALP with the plasma membrane or an induced contraction of the cell bodies.

The preceding examples showed that impedance measurements of cell-covered gold electrodes in which the cells serve as sensory elements can be used in screening assays for pharmaceutical compounds, but also for cytotoxicity screening. The interested reader is referred to Ref. 26 and 27.

SUMMARY AND OUTLOOK

Impedance spectroscopy is a general technique with important applications in biomedical research and medical diagnostic practice. Many new applications are currently under investigation and development. The potential of the technique is obviously great, since it is noninvasive, easily applied, and allows on-line monitoring, while requiring low cost instrumentation. However, there are also difficulties and obstacles related to the use of IS. Foremost, there is no separate access to the individual processes and components of the biological system, only the total impedance is measured, and this signal must be interpreted by some chosen model. There are many fundamental issues yet to be solved, both connected with understanding the origin of bioimpedance, methodological problems with finding standardized ways of comparing different samples, as well as technical issues connected with the equipment used to probe bioimpedance.

Prospective future *in vivo* applications include quantification of ischemia damage during cardiac surgery (28) and organ transplantation (29), as well as graft rejection monitoring (30). Impedance spectroscopy are also used for tissue characterization, and recently a device for breast cancer screening became commercially available. Multifrequency electrical impedance tomography (EIT) performing spatially resolved IS is a potential candidate for diagnostic imaging devices (31), but due to poor resolution power compared to conventional methods like MR, only few clinical applications are described.

The use of impedimetric biosensor techniques for *in vitro* monitoring of cell and tissue culture is promising. With these methods, high sensitivity measurements of cell reactions in response to various stimuli have been realized, and monitoring of physiological–pathological events is possible without use of marker substances. The potential applications cover pharmaceutical screening, monitoring of toxic agents, and functional monitoring of food additives. Microelectrode-based IS is interesting also for scientific reasons since it allows studying the interface between cells and technical transducers and supports the development of implants and new sensor devices (32).

Finally, affinity-based impedimetric biosensors represent an interesting and active research field (33) with many potential applications, for example, immunosensors monitoring impedance changes in response to antibody–antigen reactions taking place on electrode surfaces.

BIBLIOGRAPHY

1. Macdonald JR. Impedance Spectroscopy. New York: John Wiley & Sons; 1987.
2. Fricke H, Morse S. The electrical capacity of tumors of the breast. J Cancer Res 1926;10:340–376.
3. Schwan H. Mechanisms Responsible for Electrical Properties of Tissues and Cell Suspensions. Med Prog Technol 1993;19: 163–165.
4. Grimnes S, Martinsen ØG. Bioimpedance and Bioelectricity basics. Cornwall: Academic Press; 2000.
5. McAdams E, Lackermeier A, McLaughlin J, Macken D, Jossinet J. The linear and non-linear electrical properties of the electrode-electrolyte-interface. Biosens Bioelectron 1995;10: 67–74.
6. Kottra G, Fromter E. Rapid determination of intraepithelial resistance barriers by alternating current spectroscopy. II. Test of model circuits and quantification of results. Pflugers Arch 1984;402:421–432.
7. Cole K, Cole R. Dispersion and adsorption in dielectrics. I.. alternating current characteristics. J Chem Phys 1941;9: 341–351.
8. Cole Ks. Membrane, Ions and Impulses. Berkeley (CA): University of California Press; 1972.
9. Kell D. Biosensor. Fundamentals and Applications. Turner A, Karube I, Wilson G, editors. Oxford Science Publications; 1987. pp 427–468.
10. Schwan H. Electrical properties of tissue and cell suspensions. Advances in biological and medical physics. Lawrence J, Tobias C, editors. New York: Academic Press; 1957. pp 147–209.
11. Gersing E. Impedance spectroscopy on living tissues for determination of the state of organs. Bioelectrochem Bioenenerg 1998;45:149.
12. Giaever I, Keese CR. Monitoring fibroblast behavior in tissue culture with an applied electric field. Proc Natl Acad Sci 1984;81:3761–3764.
13. Giaever I, Keese CR. A morphological biosensor for mammalian cells. Nature (London) 1993;366:591–592.
14. Applied Biophysics, Inc. 2002.
15. Connolly P, et al. Extracellular electrodes for monitoring cell cultures. IOP Publishing; 1989.
16. Wegener J, Sieber M, Galla HJ. Impedance analysis of epithelial and endothelial cell monolayers cultured on gold surfaces. J Biochem Biophys Methods 1996;76:327–330.
17. Hagedorn R, et al. Characterization of cell movement by impedance measurements on fibroblasts grown on perforated Si-membranes. Biochem Biophys Acta—Molecular Cell Res 1955;1269:221–232.
18. Giaever I, Keese C. Micromotion of mammalian cells measured electrically. Proc Natl Acad Sci USA 1991;88:7896–7900.
19. Lo CM, Keese CR, Giaever I. Impedance analysis of MDCK cells measured by electric cell-substrate impedance sensing. Biophys J 1995;69:2800–2807.
20. Giaever I, Keese CR. Use of electric fields to monitor the dynamical aspect of cell behavior in tissue culture. IEEE Trans Biomed Eng 1986;33:242–247.
21. Mitra P, Keese CR, Giaever I. Electrical measurements can be used to monitor the attachment and spreading of cells in tissue culture. BioTechniques 1991;11:504–510.
22. De Blasio BF, Laane M, Walmann T, Giaever I. Combining optical and electrical impedance techniques for quantitative measurements of confluence in MDCK-I cell cultures. BioTechniques 2004;36:650–662.
23. Zink S, Roesen P, Sackmann B, Lemoine H. Regulation of endothelial permeability by beta-adrenoceptor agonists: Contribution of beta 1- and beta 2-adrenoceptors. Biochim Biophys Acta 1993;1178:286–298.
24. Zink S, Roesen P, Lemoine H. Micro- and macrovascular endothelial cells in beta-adrenergic regulation of transendothelial permeability. Am J Physiol 1995;269:C1209–C1218.
25. Wegener J, Zink S, Roesen P, Galla H. Use of electrochemical impedance measurements to monitor beta- adrenergic stimulation of bovine aortic endothelial cells. Pflugers Arch 1999;437:925–934.
26. Arndt S, et al. Bioelectrical impedance assay to monitor changes in cell shape during apoptosis. Biosens Bioelectron 2004;19:583–594.
27. Keese C, Karra N, Dillon B, Goldberg A, Giaever I. Cell-substratum interactions as a predictor of cytotoxity. *In Vitro* Mol Toxicol 1998;11:183–191.
28. Benvenuto, et al. Impedance microprobes for myocardial ischemia monitoring. 1st Annual International IEEE-EMBS. Lyon, France; 2000. p 234–238.

29. Haemmerich D, et al. Changes in electrical resistivity of swine liver after occlusion and postmortem. Med Biol Eng Comput 2002;40:29–33.
30. Ollmar S. Noninvasive monitoring of transplanted kidneys by impedance spectroscopy—a pilot study. Med Biol Eng Comput 1997;35:1–336.
31. Brown B. Electrical impedance tomography (EIT): A review. J Med Eng Technol 2003;27:97–108.
32. Borkholder D. Cell based biosensors using microelectrodes. Ph.D. Dissertation. 1998. Stanford University.
33. Katz E, Wilner I. Probing biomolecular interactions at conducting and semiconducting surfaces by impedance spectroscopy: routes to impedimetric immunosensors. Electroanalysis 2003;15:913–947.

See also Capacitive microsensors for biomedical applications.

IMPLANT, COCHLEAR. See Cochlear prostheses.

INCUBATORS, INFANTS

Robert White
Memorial Hospital South Bend
South Bend, Indiana

INTRODUCTION

Providing newborn infants with appropriate thermal protection is known to improve their growth rates (1–3), resistance to disease (4,5) and survival (6–11). Keeping premature, sick, or otherwise high risk babies warm is particularly critical and, when their care precludes covering them with protective swaddling clothing, especially difficult. Incubators are devices used during the care of such high-risk infants and are designed with the intent of producing environmental conditions that are consistently suitable to each unique infant's particular needs. There are many different kinds of incubators that differ in detail in the way they are constructed, heated, and controlled (12–16). All provide a mattress for the infant to lie upon, surrounded by a warmed microclimate that is controlled by a logical system governing the amount of heat needed to keep the environmental temperature within a limited range. In some incubators, this microclimate is produced within a rigid walled chamber; such devices are called closed incubators. When they are heated by using a fan to force air over a metallic heating coil prior to its entry into the infant's chamber, these closed incubators are also called forced convection incubators. There also are open incubators; those have no walls and, therefore, no chamber surrounding the mattress. There is nothing delimiting the convective environment in an open device, so they need to be heated by using a radiant warmer directed to the mattress area. These devices, therefore, are commonly called open radiant incubators, radiant warmer beds, or radiant heaters.

Each of these types of incubators provides certain unique advantages. The convectively heated incubator provides a caretaker with a far easier method for controlling the humidification of the infant's microclimate, when compared to the open radiant warmer bed. Therefore, a baby under an open radiant heater loses more body fluid than does an infant within a closed convectively heated chamber (17). But conversely, a baby in an open incubator, while more complicated to care for in terms of medical fluids administration, is physically more accessible in terms of other kinds of care that sometimes are equally important to the well being of sick babies. Current "top of the line" incubators incorporate the advantages of both types, utilizing a radiant warmer bed with a removable enclosure that allows full physical access to the infant when the incubator is operated in the radiant heater mode, and better control of humidification and noise when the enclosure is placed around the baby and operated in the convectively heated mode.

An incubator, in many respects, is just a very little house sized to fit the space and functional requirements of an infant occupant. As choices must be made when conditioning the environment in any house, different options must be considered when designing the climate control system in an incubator. In the following review, some of these considerations will be explained from the perspective of how environmental manipulators affect newborn infants who are not just little human adults, but also developing individuals with special physical, physiologic, metabolic, and neurological capabilities and limitations that make them unique. In great measure incubator manufacturers have been successful in translating present day knowledge of babies and their special needs into technical solutions that make today's incubators remarkably functional. But any infant caretaker or incubator manufacturer can attest to the limitations of today's devices which, as they are approximate to our present scientific knowledge and the existing level of technology, are flawed by our considerable remnant ignorance and the failure of existing technology to meet certain imperative needs already known.

HISTORY

It is ancient knowledge that infants who are allowed to get cold have a greater chance of dying than do infants kept warm. Prior to the nineteenth century, keeping small babies warm meant swaddling with multiple layers of cloth, providing body contact with the mother, or placement of the infant near a warm, roaring fireplace. Such classic thermal care served lusty, healthy babies well, but was inadequate to provide for the special needs of premature or otherwise enfeebled newborns. These special needs were not met because, until the last century, there was almost no recognizable major medical or social commitment toward enhancing the survival of babies born prematurely. The infant mortality rate was high and accepted. However, in response to various politicosocial events that occurred in the late 1700s and early 1800s, the value of improving premature infant survival increased, stimulating the development of special incubators in which to care for these fragile, newly valued babies.

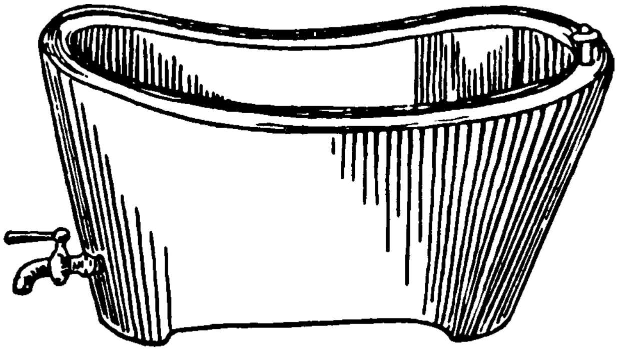


Figure 1. von Ruehl warming tub (1835). Reprinted with permission from T. E. Cone, Jr., *History of the Care and Feeding of the Premature Infant,* Boston, MA: Little, Brown, 1985.

The first serious attempt to improve on the thermal protection provided newborns was reflected in a warming tub developed in 1835 by Johann Georg von Ruehl (1769–1846) in Russia (Fig. 1). The von Ruehl tub was simply a double-walled sheet-iron open cradle that was kept warm by filling the space between the walls with warm water. Variations on von Ruehl's design were subsequently developed throughout Europe, and this type of primitive open incubator remained a standard device for care until 1878.

Although the von Ruehl device must be recognized as a developmental milestone, to be truly accurate, the developmental history of modern infant incubators must be traced back centuries to Egypt where the artificial incubation of eggs was refined and remained a closely guarded secret and uniquely Egyptian profession. Not until 1799 were these secrets introduced into Europe by members of Napoleon's expedition. Professor Stephane Tarnier (1828–1897) of the Paris Maternity Hospital in 1878 saw a chicken incubator at the Paris Zoo. The incubator, based on old Egyptian designs, had been constructed by Odile Martin. Dr. Tarnier perceived how such a device, with modifications, could be used to keep premature infants warm. Odile Martin subsequently built the first approximation of the modern enclosed infant incubator initially used at the Paris Maternity Hospital in 1880 (Fig. 2)

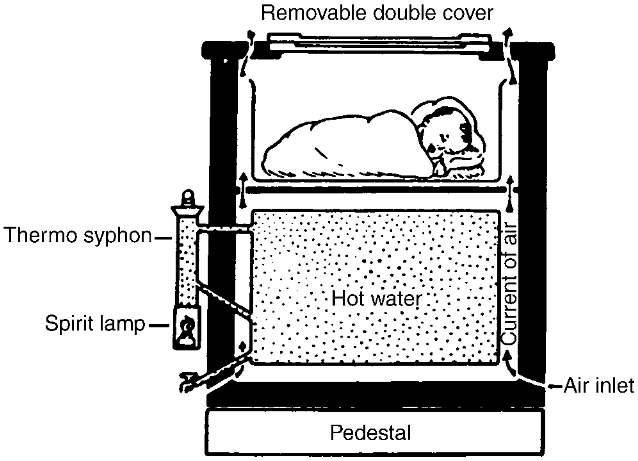


Figure 2. Tarnier incubator (1880). Reprinted with permission from T. E. Cone, Jr., *History of the Care and Feeding of the Premature Infant,* Boston, MA: Little, Brown, 1985.

The Tarnier incubator was simple in its design. The infant lay in the upper chamber of a two-chambered double-walled box insulated to slow the loss of heat. The infant chamber was topped with a removable cover through which the infant could be observed while remaining protected from cooling room drafts. The heating of the upper chamber was achieved by warming a large supply of water contained in the lower chamber of the incubator. The water was heated by an alcohol or gas lamp thermosyphon that was external to the incubator chambers and connected by piping that allowed convection driven water flow between the heater and the water reservoir. Cool room air freely flowed into the lower warming chamber where the air picked up heat from the surface of the warm water reservoir and then, by natural convection, rose to enter and warm the upper chamber containing the infant.

The Tarnier incubator was neither elegant nor efficient, and even when within the device, infants needed the extra protection of swaddling blankets. It did, however, reflect the technology of its day and, in the climate of a new commitment toward the study and welfare of feeble infants, stimulated others to refine the basic Tarnier design to correct deficiencies discovered through acquired practical experience with the incubator in clinical settings. The historical progression in this development has been illuminated by Dr. Thomas Cone, and the reader is referred to Dr. Cone's excellent treatise for a more detailed description of the many design variations introduced and tested along this early path leading to today's equipment (18).

The modernization of incubators has not only led to marked improvement in the thermal protection provided infants, but it also has been pivotal in increasing our knowledge of diseases unique to newborn babies. In turn, each increment of knowledge has required manufacturers to modify incubators in order to provide solutions to problems the new scientific discovery imposed. For example, when electric fans became available and forced air convection systems were developed, incubator heating became so improved that infants for the first time could be undressed during care. This along with the use of new transparent plastics in the construction of incubator walls allowed clinicians to make observations that led to detailed descriptions of illnesses about which little was known when infants were hidden within the predominately opaque wooden and metal chambers of the past. But while the employment of clear plastic in incubator construction enhanced the ability to observe infants, its poor insulating qualities made the task of maintaining the incubator chamber in a warm and stable state more difficult. And as improved visibility led to new lifesaving therapies, such as the administration of intravenous fluids, the use of respirators in care, and the development of new diagnostic procedures and surgical interventions, the transparent plastic walls that provided caretakers with visual access to sick babies during these therapeutic processes also served as impediments. Incubators had to be modified so that catheters, tubes, and wires could be connected to an infant's body. Increasing numbers of access holes were drilled through the walls of the incubator to provide portals of entry for these therapeutic tools, but the new fenestrations also produced new exits for life-sustaining heat. More modifications were needed and, as each problem was solved, a new dilemma emerged.

Even today, infant caretakers and incubator manufacturers continue to struggle with these and other problems contributing to the strengths and weaknesses in incubator devices. In this article, the physiologic, clinical, and technical factors leading to the design of existing incubators will be outlined further and some of the limitations of incubators explained in greater detail. Throughout, we hope that it remains clear that incubator development is an ongoing process requiring frequent and critical review of existing methods to assure that the thermal protection being provided is still appropriate during the delivery of other and especially newer forms of care also deemed necessary to a baby's well being. The ideal incubator of tomorrow is one that neither impedes care nor can itself, when providing thermal protection, be impeded by other forms of care. This has always been and remains the major challenge to health care providers and engineers committed to incubator development.

FACTORS CONSIDERED IN INCUBATOR DESIGN

Physiological Heat Balance

Even though some controversy exists concerning the exact definition of the body temperature limits within which a newborn's body functions optimally, in general, any temperature between 35.5 and 37.5 °C is probably within that normal range and can be referenced to published data available on the subject. Body temperature is determined by the balance between the heat produced within and lost from the body tissues. In order to design or even understand the design and limitations of modern incubators, a knowledge of these basic factors is required to provide the context for how infants differ from adults in their thermoregulatory capabilities.

Heat Production

All animals, including human babies, produce body heat as a by-product of the biochemical processes that sustain life. The basic amount of heat produced by a newborn infant is $\sim$1.5–2 $\text{W}\cdot\text{kg}^{-1}$. During the first weeks of life, this minimal rate of heat production is both weight and age related, with smaller and younger babies capable of producing less heat than larger and older infants (19–23).

In addition to this basic capacity, most healthy babies have the capability to generate additional heat to a maximum production rate of $\sim$4.5–5 $\text{W}\cdot\text{kg}^{-1}$ (21–23). This additional heat-producing capacity is often called upon for protective purposes, as, for example, when the infant is challenged to fight off infection or when stressed by situations that cause an exorbitant amount of heat to be lost from the body. The capability to increase the amount of heat produced to replace body heat losses is called homeothermy. In contrast to homeotherms, some creatures, such as lizards, reptiles, and fish, are poikilotherms that do not produce more heat when cooled, but actually decrease their metabolic rates when exposed to cold.

When considering thermoregulatory problems associated with newborn care, both homeothermy and poikilothermy must be understood, because under some circumstances, it is possible for a homeothermic animal to behave like a poikilotherm. Sometimes during the medical care of humans this possibility is used to a patient's benefit; for example, during some heart operations patients are given drugs that prevent their nervous systems from responding to the cold. In this circumstance, it is desirable to slow the body's metabolic rate, which can be achieved by cooling the drug treated patient who, by intent, has been changed to a temporary poikilotherm.

At times, a homeotherm may revert temporarily to a poikilothermic state. This is particularly common in immature or very sick newborns, and especially those with neurologic damage (24) or with breathing problems that lead to inadequate blood and tissue oxygen levels (25,26). Poikilothermy in a newborn can also occur because of drugs administered to a mother in labor with subsequent placental transport of those drugs to the infant (27).

In spite of their occasional reversion to poikilothermy, it nonetheless is commonly advised that newborns should be thought of as homeotherms and protected from environments that would unduly stimulate homeothermic tendencies. This is because homeotherms increase heat production by increasing metabolic work which can cause excess utilization of finite fat, sugar, and protein stores needed to sustain other vital body functions and to meet growth and developmental milestones. Moreover, extra metabolic work produces more than just extra heat; acidic and other metabolic by-products produced at the same time can cause severe imbalances in the body's critical acid–base balance (4). As a consequence of the self-protective reaction to cold stress a newborn may, therefore, be faced with an equally life-threatening biochemical stress (Fig. 3).

It has been suggested that one reason cold-exposed infants have higher mortality rates is that they become metabolically exhausted and incapable, in part because of the consequent acidosis, to fight off the stresses placed on their bodies by other threatening events. But problems can arise when attempts are made to provide infants with protection from cold stress, since it is unclear how to be sure that a given environment minimizes the infant's metabolic efforts to maintain homeothermy. Theoretically, this could be achieved by measuring the infant's metabolic rate continuously and making adjustments in the environmental supports whenever the infant's rate of metabolism changed, but measurement of heat production by a newborn is very difficult in the clinical setting.

In any case, few infant caretakers would consider the infant's metabolic rate as the only measure of success when providing a baby with thermal protection. The infant's actual body temperature is also considered important (21) and it is well known that metabolic rate is only one factor contributing to the body temperature measured. Body temperature also is influenced by the rate at which heat is lost from the body and, if one wishes to produce a device that contributes to the maintenance of an infant's body temperature, it is necessary, by virtue of its balancing

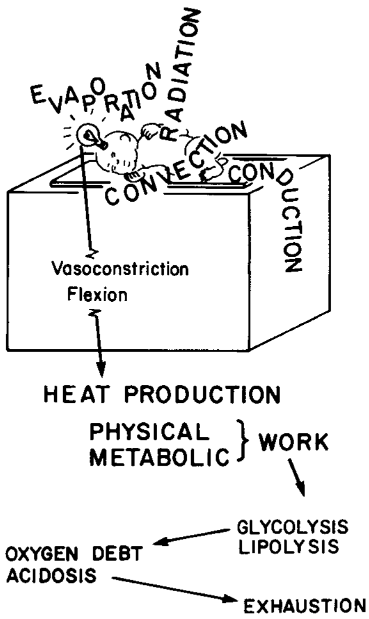


Figure 3. Schematic of homeothermy in newborns. On sensing loss of body heat, the infant minimizes heat loss from the skin by vasoconstricting blood vessels, changing body positions, and increasing metabolic rate. The increase in metabolism can produce acidosis and depletion of energy substrate stores. Reproduced by permission from P. H. Perlstein, "Routine and special care–Physical environment." In A. A. Fanaroff and R. J. Martin (Eds.), *Behrman's Neonatal-Perinatal Medicine*, 3rd ed., St. Louis, MO: C. V. Mosby Co., 1983.

effect on the final body temperature achieved, to understand how heat loss occurs in newborns.

Heat Loss

In the final analysis, incubators actually protect infants only by modifying the factors that contribute to heat loss via the well-known conductive, convective, radiant, and evaporative mechanisms.

The flow of heat occurs only when there is a difference in the temperatures of adjacent structures. Heat can only be lost by a warmer object to a cooler one. Conductive heat losses occur when an infant comes in physical contact with a cooler solid surface. A baby loses heat by conduction to a cooler mattress, blanket, diaper, or other clothing. Convective losses are similar to, but independent of, conductive losses and occur when a baby is exposed to air currents that are cooler than the infant. Convective losses are influenced not only by temperature differences, but also by the wind chill factor determined by speed at which the air is flowing. In a practical sense, this means that if an incubator has a fan as part of a forced convection heating system, air movement produced by the fan can cause a cooling effect in excess that which would occur if air at the same temperature was still.

Heat loss in infants also occurs by radiation in the infrared (IR) spectrum to cooler solid objects surrounding, but not in contact with their skin. The incubator and room walls, windows, and furniture all contribute to heat loss via radiant mechanisms. Finally, evaporative heat losses occur as infants transpire water from their skin into the surrounding environment. They can also lose heat by evaporation as residual amniotic fluid or bath water dries from their skin and hair, and they lose heat from their lungs as they exhale warm humid air.

Gross estimates of the magnitude of heat loss can be calculated by physical heat-transfer equations for each mechanism. The reader is referred to thermal transfer books for the details of these mechanisms, but examination of these equations here in a simplified form is a useful way to discover some of the special features that influence heat transfer as applied to newborn care.

All nonevaporative heat losses are quantitatively proportional to the magnitude of the temperature difference between the warmer object (T_o) losing heat and the cooler environmental feature (T_e) that will receive the heat in transfer:

$$\text{Heat loss} = \propto (T_0 - T_e)$$

This equation becomes equality by adding to it an object specific constant called a thermal transfer coefficient (k):

$$\text{Heat loss} = k(T_0 - T_e)$$

Different materials at the same temperature lose heat at different rates when exposed at the same thermal environment; for example, a block of wood has a lower thermal transfer coefficient than a block of steel. Newborn infants have higher thermal transfer coefficients than do adults and therefore lose body heat more rapidly than adults when exposed to any environment that is cooler than body temperature, so in a room temperature that feels comfortable to an adult, a newborn can get cold.

It must be emphasized that the heat loss equation as written above is grossly simplified and omits numerous other factors important to actual heat exchange. A more accurate equation would have to include a factor accounting for the infant's exposed surface area, which is a quantity that changes with changes in an infant's position and is modified if the infant is swaddled in blankets, wears a diaper, booties, hat, or, if in the course of surgical care, has a bandage applied. The equation as simplified particularly fails to reflect the true degree of complexity describing the thermal relationship between an infant's skin and the radiant surfaces of a room or incubator. The relationship is modified, for example, by complex factors describing the exact path and distance traveled by infrared (IR) waves in their transfer from object to object.

Radiant heat loss is also modified by the emissivities of the objects exchanging energy. Like black carbon, an infant's skin, no matter what its actual color, is presumed to have an emissivity of 1, which means it absorbs and emits IR rays perfectly and completely. The emissivities of the materials used in incubator manufacture or in nursery wall coverings also modify the amount of radiant exchange occurring with the infant's radiant surface. The emissivities of these objects become particularly important in an incubator chamber in which the surface area of the interior walls surrounds the exposed surface area of the infant's radiating skin.

The following equation, although still simplified, provides a better approximation of expected radiant losses (H_r)

from an infant in an incubator (28). In this equation, A_b is the exposed surface area of the infant, A_r is the area of the walls surrounding the infant, E_s is the emissivity of the infant's skin, and E_r is the emissivity of the walls. The symbol o is the Stefan–Boltzmann constant, 5.67×10^{-8} $W^{-1} \cdot m^{-2} \cdot K^{-4}$. When using this equation, temperatures are expressed in kelvin and the heat loss in watts.

$$H_r = A_b \left[\frac{1}{Es} + \frac{A_b}{A_r} \left(\frac{1}{E_r} - 1 \right) \right]^{-1} \sigma(T_s^4 - T_r^4)$$

Radiant exchange relationships are so variable because of differences between infants and different incubator environments that, even using this more complex equation, only poor and static quantitative approximation of actual radiant flux can be made in clinical settings. This proves to be a practical problem when considering incubator designs, since it has been documented that in many situations radiant losses can account for >60% of the total heat loss from an infant (29).

Evaporative heat losses are not specifically related to temperature difference and occur instead because of differences that exist between the partial pressures of water in the boundary layer of air next to the infant's skin and that in the environment beyond the boundary layer limits.

$$\text{Evap loss} = K(\text{partial pressure skin} \\ - \text{partial pressure air})(\text{ares})$$

Partial pressures are not the same as relative humidities, so even in an environment at 100% relative humidity, an infant can lose water and heat if the skin surface is warmer than the environment. For each milliliter of fluid lost from the body, ~580 g·cal (2.4 kJ) of heat are lost in the vaporization process. This route of heat loss accounts for ~25% of the total heat loss when an infant is dry. When lying unclothed on an open bed heated only by a radiant heater, up to 300 $mL^{-1} \cdot kg^{-1} \cdot day^{-1}$ of fluid can be lost by evaporation from the skin of very immature infants in the first days of life. In an enclosed incubator that is poorly humidified, up to 150 $mL^{-1} \cdot kg^{-1} \cdot day^{-1}$ of water can be lost by this mechanism in very immature infants. Following birth when the infant is wet with amniotic fluid, or following a bath, this can become the predominant route of heat loss (30–34,34).

Environmental Temperature

It should be obvious from the previous discussion that since conduction, convection, radiation, and evaporation are each relatively independent mechanisms, no single measurable quantity can be used to calculate their combined contribution to heat loss. Air temperature, for example, can be used to estimate only the convective component of heat loss from a baby, while measurements of incubator inside wall temperatures can only be helpful in determining approximate losses due to radiation. This means that if the incubator walls are cold, a baby in an incubator can lose heat even if the air temperature is warmer than the infant. The only feature necessary for this to be true is for radiant losses to be higher than convective heat gains.

Environmental temperature, although frequently used loosely to describe any ambient thermal value, must be understood to be a specific reference to the combination of temperatures actually experienced by an infant in thermal exchange relationships via multiple mechanisms. Unfortunately, few guidelines exist at present to help caretakers know the true environmental temperature for an infant within an incubator in a clinical setting. When certain conditions are met, however, some of the guidelines seem to be useful. Dr. Hey, for example, determined that in a well-humidified enclosed convectively heated incubator with single-layer Plexiglas walls, the environmental temperature perceived by a contained infant is ~1 °C lower than the measured midincubator chamber air temperature for every 7 °C difference that exists between the incubator air temperature and the air temperature of the room within which the incubator stands (35).

Heat Transfer within the Body

Since the skin of the newborn is the major heat-losing surface in exchange with the environment, mechanisms by which heat transfers from interior tissues to the skin play an important part in the heat loss process. The rate at which internally produced heat is transferred from the body core temperature T_B through body tissues to the outer body skin surface at temperatures T_s is computed using the following equation:

$$\text{Heat transfer} = C(T_B - T_s)$$

Where C is an individual's specific thermal conductance coefficient, which is affected by the absolute thickness and character of the skin, subcutaneous fat, and other subcutaneous tissue, and by the blood flow rate from the body core to its surface. Obviously, babies are smaller and have thinner body coverings than do adults, and, therefore, as they lose heat more rapidly from their surfaces than do adults, they also transfer heat more rapidly to their surfaces from their internal heat-producing organs. In addition, an infant's blood vessels are relatively close to the body surface. Since the vascularity of a particular body surface determines the rate at which blood will shunt core heat around intervening insulating tissues to the skin surface, such shunting contributes heavily to the high thermal conductance of a baby.

Heat can also be lost rapidly from an infant's body core via the respiratory system. This route of loss is of relatively minor significance in a healthy and spontaneous breathing infant, but in a baby who is ill and especially one being mechanically assisted by a respirator, this can become the major route by which body heat is lost or gained. Body heat transferred by this route is dependent on the infant's temperature, breathing rate, the flow rate and temperature of gases reaching the lungs, and the water content of the gas delivered to the airway. If temperatures and humidification are not properly monitored and controlled, the heat losses from the respiratory passages can be so great that they exceed the capacity of the incubator heater.

The Concept of a Neutral Thermal Environment

Theoretically and as demonstrated by several authors (21,36,37), it is possible for a competent homeothermic baby to have a body temperature that is below the normal range at a time when the measured metabolic rate of the infant is at a minimal unstimulated level. For example, Brück (36) documented that during a period of cooling associated with a falling environmental temperature, an infant's metabolic rate and heat production increased, but, as soon as the cooling environment was caused to reheat, the infant's metabolic rate decreased to minimal heat-producing levels, and this decrease occurred even before the infant's cooled body temperature returned to normal. This was confirmed by Adamsons et al. (21) and again in the study by Grausz (37).

The study by Adamsons et al. in particular provided some insight into why homeothermic reactions are not predicted only by examination of static body temperatures. In this study, the metabolic rates of infants in various thermal environments were determined and correlations computed to determine the relative value of measuring only rectal temperature, skin temperature, or environmental temperature, or only the difference between the skin and environmental temperatures, in reliably predicting what the infant's metabolic rates actually were at the time the temperature measurements were made. The study determined that no correlation existed between rectal temperature and metabolic rate, a slightly better correlation existed between environmental temperature and metabolic rates, a still better correlation with skin temperature existed, and an almost perfect correlation was demonstrated between metabolic rate and the difference between skin and incubator environmental temperatures (Fig. 4). These results can be understood by recalling that when body temperature is stable, heat production or metabolic rate must be equal to the rate of heat loss from the infant. If this balance does not exist, the infant will get either warmer or cooler, depending on the direction of the imbalance. So if heat production equals heat loss and heat loss is proportional only to the difference between the magnitude of the skin and environmental temperatures and not to the magnitudes themselves, it follows that heat production must also be related to the same temperature difference and, similarly, should be relatively independent of any single absolute temperature value.

These discoveries led to an approximate definition of what might constitute an optimal thermal environment in which to raise small infants. This environment is called a neutral thermal environment, referring to that set of thermal conditions existing when an infant is in a minimal metabolic state and has a body temperature that is within a normal range.

From the previous discussion, it might seem reasonable that to provide an infant with a neutral thermal environment it is necessary then only to establish skin-to-environment temperature gradients documented in various published studies to be associated with minimal metabolic rates in infants with normal temperature. Unfortunately, such references can provide only very rough guidelines, since any one infant can achieve different minimal rates of metabolism at different times and, if body temperature is to be kept stable, each change in this minimal achievable heat production rate must be balanced with a change in the amount of heat loss allowed. When the minimal heat production increases, the skin environmental temperature difference needs to be increased, and when the minimal heat production falls, the gradient needs to be decreased. Although concepts such as the neutral thermal environment can be discussed using static equations, it must be remembered that they actually are used only in dynamic settings.

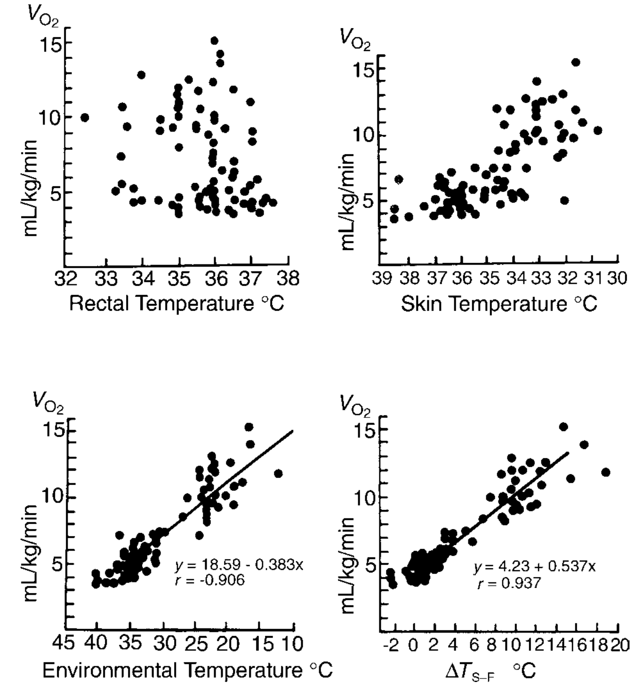


Figure 4. Metabolic rate expressed as oxygen consumption (Vo_2) as correlated with rectal temperatures, skin temperature, incubator environmental temperature, or the difference between the skin and environmental temperature (ΔT_{s-e}). Reproduced by permission from P. H. Perlstein, "Routine and special care—Physical environment." In A. A. Fanaroff and R. J. Martin (Eds.), *Behrman's Neonatal-Perinatal Medicine,* 3rd ed., St. Louis, MO: C. V. Mosby Co., 1983. Adapted from Adamsons et al. (21).

In any case, it is very difficult to provide an infant with truly neutral thermal conditions and becomes practically impossible in some common situations, such as when head hoods or other auxiliary gas delivery devices are used during care. Head hoods are small plastic boxes or tents made to enclose the infant's head when resting on a mattress. The hoods are used to deliver and control the concentration of humidified oxygen to the infant. Since the head of an infant can represent 20% of the infant's body surface, a significant amount of body heat can be lost if the head if the head hood temperature is not carefully controlled. Even if the temperature is controlled by prewarming the oxygen prior to its delivery into the head hood, the infant can lose heat if the gas is delivered at a flow rate that produces an excessive wind chill component to the convective heat flux. It also has been documented that even when total body heat losses are less than usually needed to stimulate a homeothermic response, any local cooling of facial skin tissue can stimulate a baby to become hypermetabolic (36,38).

Since no studies have been published to provide guidelines for establishing neutral thermal conditions in an incubator containing auxiliary sources of heating and cooling, and because such sources are very commonly used during infant care, no incubator manufacturer can guarantee, when an auxiliary devices are used, that such conditions can be produced by any incubator on the market today. As a corollary, unless both body temperatures and infant metabolic rates are continuously monitored, infant caretakers and medical researchers are similarly constrained from claiming precision in their delivery of continuous and certifiable thermal protection that is consistent with the concept of thermoneutrality.

Before we leave this subject, it should also be noted that a significant number of knowledgeable baby care specialists disagree with the idea that a neutral thermal environment represents an optimal goal for incubator control. Their arguments are numerous, but most often include the irrefutable fact that no one has ever documented scientifically that such protection is really beneficial. They also cite the studies by Glass et al. (2,3) in which it was documented that babies tend to lose their very important self-protective ability to react as homeotherms if not exposed to periodic cold stresses. This means that one price paid by an infant raised in a neutral thermal environment is adaptation to that environment and, much as a prolonged stay in the tropics diminishes an adult's capacity to tolerate the northern winters, a baby so adapted may be more susceptible to the damaging effects of unavoidable occasional exposures to the cold.

It must be emphasized that the arguments against the neutral thermal environment are not arguments in favor of letting all babies stay cold; the debate primarily concerns whether a baby is better off in an environment that theoretically maximizes growth by reducing metabolic work to an absolute minimum but might increase the infant's susceptibility to subsequent stresses, or better off in an environment that very mildly stimulates the infant to metabolically contribute to his own ongoing thermal welfare so that important self-protective capabilities are not forgotten. As yet there are insufficient scientific data to resolve this issue. A more recent observation is that the body temperatures of the fetus and older infant are higher than the typical neutral thermal environment proposed for preterm infants, and in both cases, follow a circadian rhythm that is not observed or supported in the typical infant incubator. These considerations imply that while the "neutral thermal environment" is a useful concept for current incubator design, it is not yet known how the "optimal thermal environment" should be defined, especially for the preterm infant.

INCUBATOR STUDIES

In spite of the difficulties encountered when attempting to define, let alone achieve, the optimal environmental conditions that are protective for small babies, it is clear that babies raised in different environments do have different survival rates. The scientific studies documenting these differences in survival in different environments are worth reviewing, since they have provided insight into features distinguishing some incubators from others and ways in which these features may produce environments that can prove to be both protective to some infants and dangerous for others. These studies have also been the major impetus to changes in designs that have resulted in the kinds of incubator devices in use today.

With few exceptions, until the early 1970s, incubator designers relied only on convective heaters to warm the chamber within which a baby was contained. Such devices were relatively simple to construct and provided a method whereby the chamber mattress could be kept warm thereby limiting conductive heat losses, and a method to keep the surrounding air warm, limiting convective losses. The use of wet sponges early in the history, and later evaporation pans, over which the convective currents of warmed air passed before entering the infant's chamber, provided the humidity needed to limit evaporative losses. Additionally, the warmed air contained in the chamber produced some warming of the chamber walls thereby reducing to some degree radiant heat losses from the infant. The heating units in early models of these incubators were controlled only by simple air temperature-sensitive thermostat mechanisms.

Such an incubator with clear plastic walls for enhancing visualization of the contained infant was used in a famous series of infant survival studies published between 1957 and 1965. In this incubator, a fan was used to force the convective air currents into the infant's chamber after passage over a heating element through a turbulence producing baffle resting in a humidifying pan of water. A highlight of these studies was published in 1958 by Silverman et al. (7) who compared the survival rates of two groups of premature infants cared for in this convectively heated and humidified device. For one group of infants the incubator air was heated to 28 °C and for the other group the air was heated to a warmer 32 °C. The study resulted in a conclusion that infants cared for in the warmer incubator had a higher survival rate than did infants cared for in the cooler incubator. During the study it was observed that although babies in the warmer incubator had a greater chance of surviving, not all of the infants who survived in the warmer incubator had warm body temperatures; in fact, 10% of the babies in the 32 °C incubator had body temperatures ~35.5 °C. Dr. Silverman deduced that the reason some of the babies got cold was due to excessive radiant heat losses to the thin plastic chamber walls that were cooled by virtue of their exterior surfaces being exposed to the cooler nursery environment.

Dr. Silverman wished to test whether even better survival rates could be achieved by assuring that an infant was not only in a warm incubator, but that the infant's body temperature also was always kept within a normal range. Along with Dr. Agate and an incubator manufacturer he helped develop a new incubator that was radiantly heated to reduce the radiant losses observed when the incubator was only convectively heated (39). To assure that the contained infant's temperature was maintained within normal range, the new incubator employed an electronic feedback servo-control mechanism that responded to changes in the temperature of a thermistor attached to

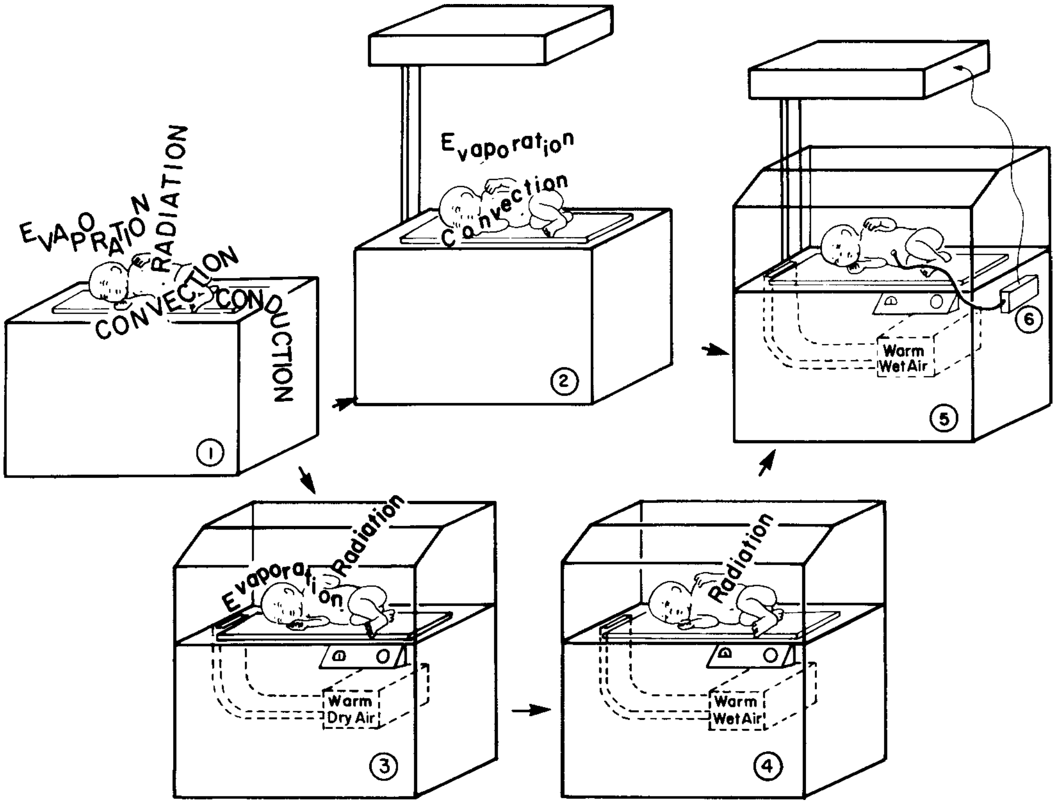


Figure 5. Logic leading to development of skin servo-controlled radiantly heated convectively ventilated incubator. (1) An unprotected baby loses heat from skin surfaces by conduction, convection, evaporation, and radiation. (2) A radiant heater eliminates radiant and conductive losses, but not those caused by convection and evaporation. (3) An unhumidified convectively heated incubator eliminates convective and conductive losses, but not those caused by radiation and evaporation. (4) Humidifying a convectively heated incubator eliminates all major losses except for the losses by radiation. (5) Using a radiant heater to warm a convectively ventilated and humidified incubator should eliminate all sources of heat loss from the infant's skin. (6) Normal infant temperature can be ensured by adding a controller to the incubator so that power is delivered to the radiant heater whenever the infant's skin temperature falls below a preset value. Reproduced by permission from P. H. Perlstein, "Routine and special care—Physical environment." In A. A. Fanaroff and R. J. Martin (Eds.), *Behrman's Neonatal-Perinatal Medicine,* 3rd ed., St. Louis, MO: C. V. Mosby Co., 1983.

the infant's skin surface, causing the incubator's radiant heater to turn on or off, depending on whether the transduced skin temperature value was below or above an absolute temperature value considered normal (Fig. 5).

Note that before settling on a servo-controlled, radiantly heated and convectively ventilated system Agate and Silverman did explore alternative methods whereby an incubator could be equipped to guarantee that an infant's temperature was maintained within a normal range. In particular, they considered simply using the well-established convective heating system in the servo-control loop but rejected this approach when they discovered that when servo controlled in response to changes in skin temperature, the convective heater produced massive and unacceptable changes in air temperature within the incubator chamber. The servo-controlled radiant heater, however, produced an environment in which the air temperature was quite stable, especially when compared to the thermal cycling recorded within the convectively heated servo-controlled system.

The radiantly heated, convectively ventilated, and skin servo-controlled enclosed incubator was evaluated in two independent studies, with results published in 1964 (9,10). In these controlled trials, premature infants were divided into two groups: one group of infants was provided care in the new radiantly heated incubator that was servo controlled to maintain the contained infant's skin temperature at 36 °C, while the other group of like babies was cared for using the simpler 32 °C air temperature thermostat-controlled convectively heated incubator that Silverman's group concluded was the best incubator setup in their previous study published in 1958. The two studies in 1964 reached a common conclusion; the skin temperature-controlled radiantly heated system produced higher survival rates when used during the care of the infants studied. Because of fabricating difficulties, though, this radiantly heated incubator model was commercially marketed for only a short period of time; soon after its introduction, it was replaced on the commercial market by a skin servo-controlled convectively heated enclosed incubator that was easier to fabricate and, like the radiantly heated device, was also capable of keeping an infant's skin temperature at a value considered normal.

The introduction of this convectively heated servo-controlled device on the market was justified by an extrapolated interpretation of the studies reported in 1964. This common interpretation led to a conclusion that the studies simply demonstrated that, in terms of survival, it was only important to keep a baby's skin temperature warm and stable. The interpretation ignored the fact that more than just a difference in skin temperatures distinguished the two study groups. Infants in the two different study environments, the one produced by an air temperature referenced thermostat controlling a convective heater and the other by a skin temperature referenced servo system controlling a radiant heater, were also, as previously well discussed by Agate and Silverman, exposed to environments that differed in the frequency and amplitude of thermal cycling produced by the different systems (39). The radiantly protected infants, who survived better, not only had warmer and more stable skin temperatures as a group, but were also exposed to a much more stable environment than were the convectively protected infants with the less favorable group outcomes.

The commercially released convectively heated and skin temperature referenced servo-controlled incubator was the most common incubator in clinical use during the late 1960s; not until 1970 was the characteristic rapidly changing air temperatures within the incubator chamber redescribed and shown to cause some sick small babies to stop breathing (40). These episodes of respiratory arrest, called apneic spells, were specifically observed during the incubator's heating cycles. The mechanism whereby a sudden rise in temperature causes some babies to become apneic remains unknown, but was an observation well reported even prior to the 1970 publication. Even without knowing the mechanism of this relationship, it remains undisputed, so incubator manufacturers have continued to search for ways to produce stabilization of the incubator environmental temperatures (Fig. 6).

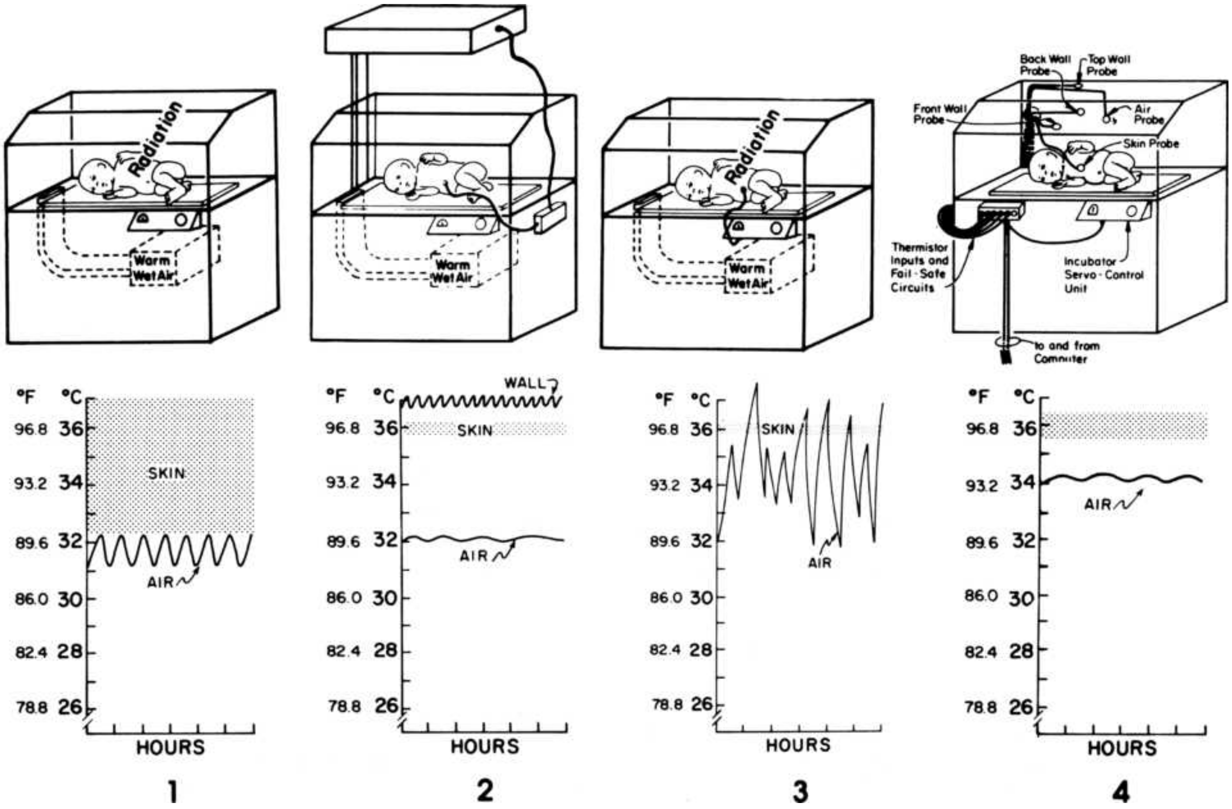


Figure 6. Skin and air temperature characteristics recorded using four different incubator systems.(1) A convectively heated and humidified incubator in which air temperature is thermostatically controlled. This was the device studied by Silverman in 1958 (7). Note cyclie variations in air temperature and wide variation in recorded skin temperatures.(2) A radiantly heated convectively ventilated and humidified incubator that is servo controlled to maintain skin temperature at specified value. This was the device studied by Day (9) and Beutow (10) in 1964. Note that the walls are warm, limiting radiant heat losses and that air temperature is stable and skin temperature variations are minimal. (3) A convectively heated and humidified incubator in which the air temperature heating is servo controlled to maintain skin temperature at specified value. This was the device reported to cause some babies to stop breathing (40) as a response to erratic heating cycles that produces rapid increases in air temperature. (4) A convectively heated and humidified incubator that is computer controlled using Alcyon algorithm (11). Note the stable air temperature and minimal variability in skin temperature.

INCUBATOR DYNAMICS

There are many reasons why attempts to stabilize incubator heating have been only partially successful. It is fairly simple to build a box within which the temperatures can be predictably controlled and stabilized for prolonged periods of time if the box is never opened and the thermal characteristics of both the box and its contents never change, but these simplifying conditions never exist when caring for a sick infant. When infants are cleaned, fed, examined, or otherwise cared for, they must be touched, and the incubator box must be entered. Infant's body positions frequently change, exposing different surface areas with different shapes to the incubator environment causing changes in convective flow patterns and altering the view factors influencing radiant heat flow to the incubator walls. Incubator openings necessitated by the need to touch a contained infant cause both environmental cooling and increased infant heat loss that, in an incubator heated in response to either air temperature or infant temperature changes, causes a logical increase in the incubator's heat output. If any such heating requirement is sustained, the incubator heating element warms to the point where it will retain and release heat even after the incubator is closed and the need for additional heating has passed. Such heat retention and release contributes to what is commonly referred to as the thermal lag characteristic of a heating system and can cause temperatures to overshoot, that is to rise above the temperature level targeted when the heater was activated. This is the same phenomenon observed when an electric stove is turned off, and yet the heating element continues to glow brightly prior to cooling. As with an electric stove heating element, the heater in an incubator is not only slow to cool, but also relatively slow to warm up when first energized after the heater is turned on. Again, just as unintended overheating can occur, the thermal lag due to the mass of the heater can result in an incubator getting colder than intended because of this characteristic delay between action and reaction.

Besides the heater element, there are numerous other places where heat is stored in the incubator system. For example, heat storage occurs in the water used for humidification and in the air masses between the heater and the incubator chamber. Since these must be heated to a temperature higher than the incubator chamber in order to raise the chamber temperature, the heat stored in these parts also will continue to raise the chamber temperatures even after the heater power supply has been turned off. Conversely, when the heater power is turned back on, not only the heater but also the air in the spaces leading to the chamber must heat up before the temperature of the air in the chamber can rise.

It is these delays between the time the heater power is changed and the time the air temperature responds that determines the magnitude and frequency of the air temperature cycles to which an incubated infant is exposed. Thermal lag obviously contributes to the tendency for incubator environments to become unstable and, thereby, potentially threatening to the contained infant. Many hardware and logical software solutions to this problem have been tried in commercially available devices, but all have been frustrated by the complex nature of newborn care, which results in a degree of unpredictability beyond the compensating capability of any solution thus far tried. The implementation of feedback control on incubator heating is an example of one way to attempt to respond to many of these problems. However, examining how servo mechanisms actually work in a little more detail provides a deeper appreciation of why this logical technology often fails in the dynamic setting of an incubator and may even contribute to instability in the incubator environment.

Feedback Control

Feedback control systems are commonly referred to as closed loop control systems, as contrasted to open loop systems. A cooking stove again can be used to give an example of each type of control system. Stove top heaters are typically controlled by an open loop system. That is, a dial is adjusted controlling the quantity of gas or electricity going to the heater unit. In this manner, a fixed rate of heat production by the heater unit is specified. This is called an open-loop control system because after once setting the rate of heat production, the heater will continue to produce the same heat output regardless of how hot the object on the stove gets.

In contrast, the oven of a modern stove is equipped with a closed-loop temperature control system, in which a dial is adjusted to specify the desired oven temperature. In control system parlance, this specified temperature is referred to as the set point. A temperature sensor inside the oven works in conjunction with the temperature setting dial to control the rate of heat production in the oven heating unit and when the temperature measured by the oven temperature sensor rises to the set point value on the control dial, the oven heat production is reduced or stopped entirely. After the heat production is stopped, the oven temperature slowly falls as the heat escaped from the oven to the surrounding area. At some point, the oven temperature will fall below the temperature set on the control and the heater will again be turned on; this on–off cycling will continue as long as the oven is in operation.

Feedback Control and Incubators

An incubator heated in automatic feedback response to changes in electronically transduced infant temperature is called an infant skin servo-controlled (ISC) incubator. In one type of ISC incubator the heater is instructed to turn completely on when the infant's skin temperature falls below a preset lower limit or turn completely off when skin temperature exceeds a defined upper limit. Because power applied to the heater is either maximal or zero, this form of servo mechanism is called nonlinear or "on–off control". Another form of control is designated as linear proportional servo control. In a proportional control system the amount of power applied to the incubator heater is graduated in a manner to be proportional in some linear fashion to the transduced skin temperature deviation from a predetermined value. The amount of power actually applied to the heater for each incremental change in skin temperature can be different in different realizations of this control method, and this increment determines the amount of deviation in skin temperature that can occur before full or zero power is applied.

The theoretical advantage of a proportional over an on–off servo-control system is the possibility of limiting the tendency that a large heating element has to overshoot or undershoot a desired heater temperature. In a proportional system, the heater is turned off slowly as the infant's skin temperature warms toward the set point temperature. This contrasts with an on–off system that remains fully energized until the set point is reached. In an incubator system, a properly designed skin temperature referenced proportional control unit will produce very stable environmental temperatures as long as the infant is stable, the temperature sensing the thermistor attached to the skin remains undisturbed, and the incubator chamber is kept closed and protected from outside environmental perturbations. In a clinical setting such stable and undisturbed conditions exists infrequently, so the theoretical advantages of proportional control over on–off control are difficult to demonstrate. In fact, the cycling of temperatures recorded within a proportionally controlled incubator in a dynamic clinical setting can be indistinguishable from that recorded in an on–off controlled incubator, and this functional behavior is predicted in basic feedback control theory (Fig, 7).

The thermal cycles recorded in servo-controlled incubators are produced as a combined consequence of (*1*) an inherent characteristic of closed-loop feedback systems; (*2*) periodic extreme perturbations of the skin thermistor that trigger either full or zero energization of the incubator heater; and (3) the incubator's thermal lag characteristic, which causes repetitive over-and undershoot of temperatures targeted by the control logic. Following the initiation of a cycling pattern, the time it takes the system to settle down and reestablish stable control is variable and related to the heat-dissipating speed of the heater material and the thermal buffering capability of the homeothermic infant. In some situations, the cycling, when induced by a single perturbation, has been observed to continue for many hours and, when an incubator is repeatedly perturbated, for many days. The published evidence that some babies react to these thermal cycles by becoming apneic justifies a reminder that these characteristic thermal cycles represent a profound problem negating some of the advantages intended when feedback control is applied in incubator designs. At least in incubator servo systems that use skin temperature as a reference value to determine heater status, even the theoretical negative effects often outweigh any theoretical or demonstrable effects that are positive. This can be appreciated by recalling that a *sine qua non* of optimal control in a skin temperature referenced servo system is the reliable and accurate measurement of an infant's skin temperature and, using today's technology in a clinical setting, the ability to transduce skin temperatures accurately for prolonged periods of time is marginal at best and, perhaps, even impossible.

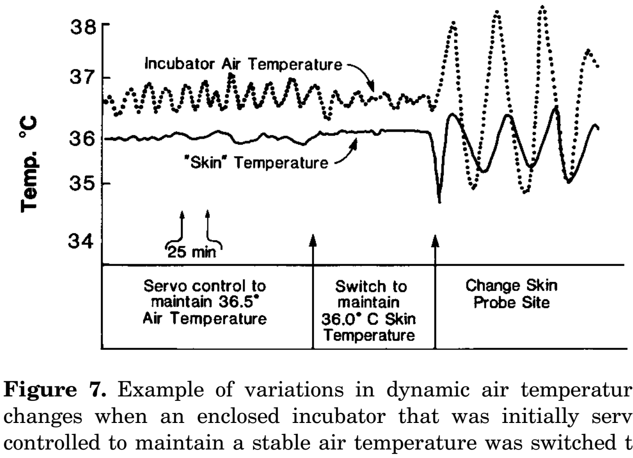


Figure 7. Example of variations in dynamic air temperature changes when an enclosed incubator that was initially servo controlled to maintain a stable air temperature was switched to a skin temperature referenced servo control mode and then was perturbated by changing the site of attachment of the skin temperature-sensing thermistor. The wide thermal cycling illustrated as a consequence of this sequence continued for 3 h before returning to a more stable pattern.

Skin Temperature Measurement

Skin temperature measurement accuracy is limited by variability in the characteristics of infant, transducers, and infant care techniques. The surface temperatures of infants are not homogeneous because of difference in (1) skin and subcutaneous tissue thickness over different body parts, (2) differences in structures underlying different skin surfaces, and (3) difference in the vascularity and vasoreactivity-characterizing different body regions. Different skin surfaces have different temperatures, and when temperature transducers are connected to the skin surface, they measure the temperature only at the specific site of their connection. Moreover, thermistors are attached using devices that can compression of superficial skin vessels underlying the thermistor element. Thus, by their very attachment, thermistors modify both the absolute temperature they are expected to measure and the spontaneous dynamic variability in the measured skin temperature that is normally affected by the changing amounts of warm blood flowing through the skin over different time periods. Additional factors also affect thermistor accuracy. Thermistors are faulted as precise skin temperature measuring devices because they are manufactured in various shapes and sizes and are protected using different materials so that each affects transduction in a specific and different way. Thermistors also generally measure temperature three dimensionally (3D). They are affected not only by the temperature of the surface to which they are attached, but also by the environment to which their unattached surfaces are exposed. Depending on the amount and type of insulation used in their manufacture, they are also affected by the temperature of the wires used to connect the thermistor to electronic signal conditioners.

These inherent characteristics of thermistors, added to the fact that they are freely moved during clinical care from one site of attachment to another, provide sufficient cause to explain why the skin temperature-dependent heater-controlling servo mechanism in an incubator can easily be directed into an unstable state that produces thermal cycling in the environment. This environmental instability is even further exacerbated by infant care practices that, for examples, cause thermistors to be removed from the skin when X rays are taken, or to be covered with sterile towels during surgical procedures. During such care-related events, the thermistor is fooled into measuring environmental and not skin temperature. Obviously, if the thermistor provides the servo electronics with misinformation, the servo control decisions based on this information can be nonsensical.

THE NONTHERMAL ENVIRONMENT

Although infant incubators were initially designed solely to maintain body temperature in high risk infants, they are now seen in a much more complex role, as devices that provide a complete "microenvironment" for high risk infants. To one extent or another, they are used to modify the sensory input an infant receives through visual, auditory, olfactory, and kinesthetic pathways. In addition, incubators are now appreciated as the source of exposure to potentially unwanted environmental toxins, such as electromagnetic radiation (EMR) and chemical compounds used in the manufacture and operation of the incubator. Closed incubators are often used as delivery systems for oxygen and humidification, sometimes in ways unanticipated at the time of their design and manufacture. Incubators have been developed for specialized purposes, such as transport, use in an magnetic resonance imaging (MRI) suite, or cobedding twin infants. Incubators have increasingly been designed as a platform for a modular system of support equipment including ventilators, IV pumps, and monitors. As these devices become increasingly controlled by digital components, they also gain the capability of integrating their data output, so that incubator-derived information, such as air temperature and the infant's skin temperature, can be continuously recorded in an electronic medical record. Incubators have had a role in infection control since their earliest days, but this function is now being increasingly emphasized as infection has emerged as the leading cause of late morbidity in premature infants. These multiple functions of modern incubators increase their complexity exponentially, since many of these factors interact with one another, often in ways unanticipated by either the designers or the users. Because of the recent nature of these nonthermal applications of the infant incubator, there is only a limited scientific foundation to guide designers and caregivers, so not all of these topics will be discussed in greater depth below.

The Infant Incubator as a Sensory Microenvironment

Although the comparison of an incubator to a uterus, the environment it replaces, has been noted since the earliest days of incubator design, it will be evident from the preceding sections of this article that temperature regulation has been the first and primary design consideration: and necessarily so, since it had immediate implications for infant survival. When incubators became used as devices for delivery of supplemental oxygen in the mid-twenteth century, morbidity and mortality were again the primary endpoints: first for improved survival as the benefits of oxygen supplementation were identified, and then for increased morbidity as an epidemic of oxygen-induced blindness from retinopathy of prematurity followed, again chronicled most eloquently by Dr. Silverman (41). Only recently have the other environmental features of the uterus been compared to the micro and macro environments of the NICU and the implications for design been considered.

Taste, smell, and touch are the earliest fetal senses to develop, beginning in the second trimester of pregnancy, followed closely by auditory development, and finally by visual development as the baby approaches term gestation. While thus far there appears to be no reason to suspect that the sense of taste is stimulated or influenced by the incubator, there is accumulating evidence that the other senses are indeed affected by this microenvironment.

Infants can develop a conditioned response to certain odors that they come to associate with painful procedures, such as alcohol, whereas a pleasant odor has been shown to reduce apnea, and babies will orient preferentially to odors from their mother (42).

In utero, infants are exposed to a fluid environment, frequent movement with circadian rhythmicity, and as they approach term, increasing contact with a firm boundary, features absent in the typical incubator. After-market adaptations that caused the bed or the entire incubator to move in one fashion or another have been introduced sporadically, often with the intent of reducing infant apnea, but none have been documented to be efficacious. Additional modifications of the infant mattress to make it more suitable for the skin and developmental needs of preterm infants have been introduced, again without clear benefit to this point.

Sound is a constant although variable stimulus in the uterus, and different in many ways from that of the modern incubator. *In utero* sounds are delivered via a fluid medium where lower pitched sounds predominate, and the mother's voice, heartbeat and bowel sounds are far more prevalent than any extraneous noise. As such, there is a definite circadian rhythm both to the sounds and to movement associated with them. In the closed incubator, the predominant sound is that of the incubator fan that produces a constant white noise, usually in excess of 50 dB, a level that has been shown to interfere with infant sleep (43) and speech recognition (44). Depending on the NICU in which it is used, this may be louder or softer than the NICU ambient noise level, so the closed incubator may be used as a haven from a noisy NICU, or itself could be noisier than the external environment, in which case the open radiant warmer might provide a more suitable auditory environment. A number of after-market modifications have been suggested to reduce both fan noise and intrusion of noise from outside the incubator, including blankets or quilts to cover the incubator and sound-absorbing panels (45,46), but their use is problematic to the degree that they affect air flow and other operating characteristics of the incubator.

The visual environment of the incubator may be presumed to be neutral, but incubators are often used to control the visual environment of the NICU, particularly in conjunction with the use of incubator covers, to produce a dimly lit environment which may enhance infant sleep (47). Light penetration into the incubator may also affect circadian rhythmicity in the preterm infant (48), and may be a source of heat gain through the greenhouse effect.

Electromagnetic Radiation in the Infant Incubator

Any electrically powered device emits electromagnetic radiation (EMR), usually at intensities far below those considered to constitute a risk. Since the organs of preterm infants are in a crucial and rapid phase of growth, however, concern about EMR emitted by incubator and radiant warmer heaters and other components has merited special attention. Several studies have documented EMR levels in incubators, with proposed strategies to reduce exposure including shielding panels (49) and increasing the distance between the baby and the EMR source (50).

Incubators for Specialized Purposes

The conventional closed incubator or radiant warmer is used as a static device in the NICU, but the same needs for temperature control and a safe microenvironment exist for infants who require transport from one hospital to another, or to an MRI suite. Transport incubators place a premium on space (so that multiple modular components can be mounted) and weight (especially those used for air transport). Incubators developed for use in an MRI suite must be similarly portable, but use almost exclusively plastic materials and have an integrated coil for scanning (51,52).

SUMMARY

Infant incubators are specially heated devices that provide a bed surface or chamber within which an infant can be cared for and kept warm. Throughout this article both the positive and negative features of today's incubators have been noted and placed into both a context of what is theoretically desirable and of what is practically feasible. It is apparent that, at present, our knowledge of infant physiology and the availability of technical solutions are severely limited, and that all existing incubator devices reflect and are faulted by these limitations. In historical perspective, however, it is clear that incubators over the past 100+ years have steadily been improved by manufacturers to incorporate new items of knowledge and technology as they become available. This historical path has been fruitful and provides, in its continuing course, direction for the future in incubator development. Future iterations of the infant incubator will need to incorporate new information on the optimal microenvironment for the high-risk newborn as well as new capabilities made possible by the ongoing quantum changes in digital technology.

BIBLIOGRAPHY

1. Meystán J, Járai I, Fekete M. The total energy expenditure and its components in premature infants maintained under different nursery and environmental conditions. Pediatr Res 1968;2:161.
2. Glass L, Silverman WA, Sinclair JC. Effect of the thermal environment on cold resistance and growth of small infants after the first week of life. Pediatrics 1968;41:1033.
3. Glass L, Silverman WA, Sinclair JC. Relationship of thermal environment and caloric intake to growth and resting metabolism in the late neonatal period. Biol Neonate 1969;14:324.
4. Gandy GM, et al. Thermal environment and acid base homeostasis in human infants during the first few hours of life. J Clin Invest 1964;43: 751.
5. Cornblath J, Schwartz R. Disorders of carbohydrate metabolism in infancy. In: Shaffer JA, editors. Major Problems in Clinical Pediatrics. Vol. 3. Philadelphia: Saunders; 1966. p 34.
6. Silverman WA, Blanc WA. The effect of humidity on survival of newly born premature infants. Pediatrics 1957;20:477.
7. Siverman WA, Ferrig JW, Berger AP. The influences of the thermal environment upon the survival of newly born premature infants. Pediatrics 1958;22:876.
8. Silverman WA, Agate FJ, Ferrig JW. A sequential trial of the nonthermal effect of atmospheric humidity on survival of newborn infants of low birth weight. Pediatrics 1964;34:171.
9. Day RL, Caliguiri L, Kamenski C, Ehrlich F. Body temperature and survival of premature infants. Pediatrics 1964;34: 171.
10. Beutow KC, Klein SW. Effect of maintenance of normal skin temperature on survival of infants of low birth weight. Pediatrics 1964;34:163.
11. Perlstein PH, Edwards NK, Atherton HD, Sutherland JM. Computer assisted newborn intensive care. Pediatrics 1976; 57:494.
12. Evaluation of infant radiant warmers. Health Devices, 1973;3:4.
13. Perstein PH. Thermal control. Rep Ross Conf Pediatr Res 1976;69:75.
14. Edwards NK. Radiant warmers. Rep Ross Conf Pediatr Res, 1976;69:79.
15. Evaluation of infant incubators. Health Devices. 1981;11:47.
16. Evaluation of infant incubators. Health Devices 1982;11:191.
17. Wu PYR, Hodgman JE. Insensible water loss in preterm infants. Pediatrics 1974;54:704.
18. Cone Jr. TE, History of the Care and Feeding of the Premature Infant. Boston: Little, Brown; 1985.
19. Scopes JW. Metabolic rate and temperature control in the human body. Br Med Bull 1966;22:88.
20. Scopes JW, Ahmed I. Minimal rates of oxygen consumption in sick and premature newborn infants. Arch Dis Child 1966;41:407.
21. Adamsons K Jr., Gandy GM, James LS. The influence of thermal factors upon oxygen consumption of newborn infants. J Pediatr 1965;66:495.
22. Hill JR, Rahimtulla KA. Heat balance and the metabolic rate of newborn babies in relation to environmental temperature: And the effect of age and weight on basal metabolic rate. J Physiol(London) 1965;180:239.
23. Dawes GS. Oxygen consumption and temperature regulation in the newborn. I Foetal and Neonatal Physiology. Chicago: Year Book Medical Publishersh; 1968. p 191.
24. Cross K, et al. Lack of temperature control in infants with abnormalities of the central nervous system. Arch Dis Child 1971;46:437.
25. Dawkins MJR, Hull D. Brown fat and the response of the newborn rabbit to cold. J Physiol (London) 1963;169:101.
26. Hill JR. Oxygen consumption of newborn and adult mammals: Its dependence on oxygen tension in inspired air and on environmental temperatures. J Physiol (London) 1959;149: 346.
27. Cree JE, Meyer J, Hailey DM. Diazepam in Labour: Its metabolism and effect on the clinical condition and thermogenesis of the newborn. Br Med J 1973;3:251.

28. Brück K. Heat production and temperature regulation. In: Stave U, editor. Pernatal Physiology. New York: Plenum Press; 1978. p 474.
29. Day RL. Respiratory metabolism in infancy and childhood. Am J Child 1943;65:376.
30. Hey EN, Katz G. Evaporative water loss in the newborn baby. J Physiol (London) 1969;200:605.
31. Sulyok E, Jéquier E, Ryser G. Effect of relative humidity on thermal balance of the newborn infant. Biol Neonate 1972; 21:210.
32. Belgaumkar TR, Scott KE. Effects of low humidity on small premature infants in servo control incubators. Biol Neonate 1975;26:337.
33. Hammarlund K, Nilsson GE, Öberg PA, Sedin G. Transepidermal water loss in newborn infants. Acta Paediatr Scand 1977;66:553.
34. Hammarlund K, Nilsson GE, Öberg PA, Sedin G. Transepidermal water loss in newborn infants: Evaporation from the skin and heat exchange during the first hours of life. Acta Paediatr Scan 1980;69:385.
35. Hey EN, Mount LE. Heat losses from babies in incubators. Arch Dis Child 1967;42:75.
36. Brück K. Temperature regulation in the newborn infant. Biol Neonate 1961;3:65.
37. Grausz JP. The effects of environmental temperature changes on the metabolic rate of newborn babies. Acta Paediatr. Scand 1968;57:98.
38. Mestán J, Jrai I, Bata G, Fekete M. The significance of facial skin temperature in the chemical heat regulation of premature infants. Biol Neonate 1964;7:243.
39. Agate FJ, Silverman WA. The control of body temperature in the small newborn infant by low-energy infra-red radiation. Pediatrics 1963;37:725.
40. Perlstein PH, Edwards NK, Sutherland JM. Apnea in premature infants and incubator air temperature changes. N Engl J Med 1970;282:461.
41. Silverman WA. The lesson of retrolental fibroplasias. Sci Am 1977;236:100.
42. Schaal B, Hummel T, Soussignan R. Olfaction in the fetal and premature infant: Functional status and clinical implications. Clin Perinatol. 2004;31:261.
43. Morris BH, Philbin MK, Bose C. Physiologic effects of sound on the newborn. J Perinatol 2000;20:S55.
44. Robertson A, Stuart A, Walker L. Transmission loss of sound into incubators: implications for voice perception by infants. J Perinatol 2001;21:236.
45. Johnson AN. Neonatal response to control of noise inside the incubator. Pediatr Nurse 2001;27:600.
46. Bellini CV, et al., Use of sound-absorbing panel to reduce noisy incubator reverberating effects. Biol Neonate 2003;84:293.
47. Hellstrom-Westas L, et al., Short-term effects of incubator covers on quiet sleep in stable premature infants. Acta Paediatr 2001;90:1004.
48. Rivkees SA. Emergence and influences of circadian rhythmicity in infants. Clin Perinatol 2004;31:217.
49. Bellieni CV, et al. Reduction of exposure of newborns and caregivers to very high electromagnetic fields produced by incubators. Med Phys 2005;32:149.
50. Bellieni CV, et al. Increasing the engine-mattress distance in neonatal incubators: A way to decrease exposure of infants to electromagnetic fields. Ital J Pediatr 2005;29:74.
51. Blumi S, et al. MR imaging of newborns by using an MR-compatible incubator with integrated radiofrequency coils: Initial experience. Radiology, 2004;231:594.
52. Whitby EH, et al. Ultrafast magnetic resonance imaging of the neonate in a magnetic resonance-compatible incubator with a built-in coil. Pediatrics 2004;113:e150.

Further Reading

Adamsons K. The role of thermal factors in fetal and neonatal life. Pediatr Clin North Am 1966;13:599.

Ahlgren EW. Environmental control of the neonate receiving intensive care. Int Anesthesiol Clin 1974;12:173.

Brück K. Heat production and temperature regulation. In: Stave U, editor. Perinatal Physiology New York: Plenum Press; 1978 p 455.

Dawes GS, Oxygen consumption and temperature regulation in the newborn. I Foetal and Neonatal Physiology. Chicago: Year Book Medical Publisher; 1968. p 191.

Delue NA. Climate and environment concepts. Clin Perinatal 1976;3:425.

Hey EN, Katz G. The optimum thermal environment for naked babies. Arch Dis Child 1970;45:328.

Holman JP. Heat Transfer, New York: McGraw-Hill; 1981.

Klaus M, Fanaroff A, Martin RJ. The physical environment. In: Klaus MH, Fanaroff AA, editors. Care of the High Risk Neonate. Philadelphia: Saunders; 1979. p 94.

Lutz L, Perlstein PH. Temperature control in newborn babies. Nurs Clin North Am 1971;6:15.

Mayr O. The Origins of Feedback Control. Cambridge, (MA): MIT Press; 1970.

Ogata K. Modern Control Engineering, Englewood Cliffs (NJ): Prentice-Hall; 1970.

Oliver TK. Temperature regulation and heat production in the newborn. Pediatr Clin North Am 1965;12:765.

Oppenheim AV, Willsky A, Young IT. Signals and Systems, Englewood Cliffs (NJ): Prentice-Hall; 1983.

Perstein PH. Thermal regulation. In: Fanaroff AA, Martin RJ, editors. Behrman's Neonatal-Perinatal Medicine, 3rd ed. St. Louis (MO): Mosby; 1983. p 259–277.

Scopes JW. Thermoregulation in the newborn. In: Avery GB, editors. Neonatology, Philadelphia: Lippincott; 1975. p 99.

Sinclair JC. The effect of the thermal environment on neonatal mortality and morbidity. In: Adamson K, Fox HA, editors. Preventability of Perinatal Injury. New York: Alan R. Liss; 1975. p 147.

Sinclair JC. Metabolic rate and temperature control. In: Smith CA, Nelson NM, editors. The Physiology of the Newborn Infant. Springfield (IL) : Thomas; 1976. p 354.

Todd JP, Ellis HB. Applied Heat Transfer. New York: Harper & Row; 1982.

See also BIOHEAT TRANSFER; NEONATAL MONITORING; TEMPERATURE MONITORING.

INFANT INCUBATORS. See INCUBATORS, INFANT.

INFORMATION SYSTEMS FOR RADIOLOGY. See RADIOLOGY INFORMATION SYSTEMS.

INFUSION PUMPS. See DRUG INFUSION SYSTEMS.

INTEGRATED CIRCUIT TEMPERATURE SENSOR

TATSUO TOGAWA
Waseda University
Saitama, Japan

INTRODUCTION

Temperature can affect the electronic characteristics of semiconductor devices. Although this is a disadvantage in many applications, especially for analogue devices, it may be turned into an advantage if such a device is used as a temperature sensor. In principle, any parameter in such a device having a temperature coefficient can be used for temperature measurement. For example, a temperature telemetry capsule, in which a blocking oscillator frequency varies with temperature, has been developed for measuring gastrointestinal temperature (1). In this system, the temperature affects the reverse-bias base-collector current, which determines the period of relaxation oscillation. However, it has been shown that the voltage across a p–n junction of a diode or transistor under a constant forward-bias current shows excellent linear temperature dependency over a wide temperature range. Many conventional or specially designed diodes or transistors composed of Ge, Si, or GaAs have been studied for use as thermometers (2–4).

The advantages of diodes and transistors as temperature sensors are their high sensitivity and low nonlinearity. The temperature sensitivity under normal operation is ca -2 mV/K, which is ~50 times higher than that of a copper-constantan thermocouple. The nonlinearity is low enough for many applications, although its value depends on the structure and material of the device. It is known that a Schottky diode, which has a structure composed of a rectifying metal-semiconductor contact, possesses good voltage–temperature linearity (5). Some transistors used as two-terminal devices by connecting the base to the collector also possess good linearity (6,7), and a transistor that has been especially developed for temperature sensing is commercially available (8). This has a linearity that is comparable to that of a platinum-resistance temperature sensor.

It is advantageous to have a diode and a transistor temperature sensor fabricated on a chip with associated interfacing electronics using integrated circuit (IC) technology. Several integrated temperature sensors that provide either analogue or digital outputs have been developed and are commercially available. A diode or transistor temperature sensor fabricated on a central processing unit (CPU) chip is especially useful when used to monitor the temperature of the chip. Such a sensor has been used to detect overheating, and to protect the CPU system by controlling a fan used to cool the chip or to slow down the clock frequency.

THEORY

The characteristics of p–n junctions are well known (9,10). In p–n junction diodes, the current flowing through the forward-biased junction is given by

$$I = I_{\rm s}(e^{qV/mkT} - 1) \tag{1}$$

where $I_{\rm s}$ is the saturation current, q is the electron charge, V is the voltage across the junction, k is the Boltzmann constant, m is the ideality factor having a value between 1 and 2, which is related to the dominant current component under the operating conditions used, and T is the absolute temperature. At a temperature close to room temperature, and when the current is relatively high, so that the current due to the diffusion of the carrier dominates, $m = 1$, and so the second term in Eq. 1 given in parentheses can be neglected. Equation 1 can then be simplified to

$$I = I_{\rm s}e^{qV/kT} \tag{2}$$

The temperature dependence of the saturation current, $I_{\rm s}$, is given by

$$I_{\rm s} = Ae^{-E_{\rm g}/kT} \tag{3}$$

where $E_{\rm g}$ is the bandgap energy at $T = 0$ K, and A is a constant dependent on the geometry and material of the device. Strictly speaking, A also depends on the temperature. However, the temperature dependency is very weak compared to the exponential term in Eq. 3. Thus,

$$I = A{\rm e}^{(qV-E_{\rm g})/kT} \tag{4}$$

For a constant current, I, $(qV–E_{\rm g})/kT$ is constant. Thus, the voltage across a p–n junction, V, is a linear function of the absolute temperature, T. On extrapolating to $T = 0$, then $qV = E_{\rm g}$.

The temperature coefficient of V can be derived from Eq. 4 as

$$\left.\frac{dV}{dT}\right|_{I=\rm const} = \frac{V - E_{\rm g}/q}{T} \tag{5}$$

Since the value of $qV–E_{\rm g}$ is always negative, V decreases with increasing T. For silicon, $E_{\rm g}$ ~1.17 eV, and for T ~300 K, V ~600 mV, and dV/dT ~-1.9 mV/K. In actual diodes, the current–voltage characteristics have been studied in detail over a wide temperature range. The forward voltage exhibits a linear dependence for $T > 40$ K for a constant current (11). The observed value of dV/dT in a typical small signal silicon p–n junction diode ranges between -1.3 and -2.4 mV/K for $I = 100$ μA (11). In germanium and gallium arsenide p–n junction diodes, and for silicon Schottky diodes, the forward voltage exhibits a similar sensitivity (3–5).

In most p–n junctions, the current through the junction contains components other than those due to carrier diffusion, and therefore, Eq. 4 does not hold. The base-emitter p–n junction in transistors is advantageous in this respect. Here, the diffusion component forms a larger fraction of the total current than that in diodes, even for a diode connection in which the base is connected to the collector. The nonlinear temperature dependence in the forward voltage in diode-connected transistors is lower than that of most diodes (7). Further improvement in linearity is attained under constant collector current operation, since only the diffusion component flows to the collector, while other components flow to the base (12).

From Eq. 2, one can obtain the following expression

$$\ln I = \ln I_{\rm s} + qV/kT \tag{6}$$

The value of T can be obtained from the gradient of a plot of ln I versus V, as q and k are known universal constants. This implies that the current–voltage characteristics can be used as an absolute thermometer (6). If ln I is a linear function of V, only two measurements are required to determine the gradient. If V_1 and V_2 are voltages corresponding to different current levels, I_1 and I_2, the difference between these two voltages is calculated using

$$V_1 - V_2 = (kT/q)\ln(I_1/I_2) \tag{7}$$

Thus, the difference in voltage corresponding to the different current levels for a constant ratio is proportional to the absolute temperature, without any offset. Using this relationship, a thermometer providing an output proportional to the absolute temperature can be realized, either by applying a square wave current to a p–n junction (12), or by using two matched devices operating at different current levels (13).

FUNDAMENTAL CIRCUITS AND DEVICES

A schematic drawing of the fundamental circuit of the thermometer with a short-circuited transistor or a diode is shown in Fig. 1. A constant current is applied to the transistor or diode in the forward bias direction, and the voltage across the junction is amplified using a differential amplifier. By adjusting the reference voltage applied to another input of the differential amplifier, an output voltage proportional to either the absolute temperature in kelvin or in degrees Celsius or any other desired scale can be obtained. The operating current of small signal diodes and transistors is typically 40–100 A. If the current becomes too high, a self-heating error may be produced due to the power dissipated in the junction. If the current becomes too small, problems due to leakage and the input current of the first stage amplifier may become significant (7).

The nonlinearity in the temperature dependency of the forward voltage is not a serious problem for most applications, and it can be reduced by appropriate circuit design. In a Schottky diode, this nonlinearity is < 0.1 K over the temperature range −65 to 50 °C (5), and a comparable performance is expected for diode-connected silicon transistors (7). Further improvement in the linearity can be attained by linearization of the circuit. Linearization using a logarithmic ratio module reduces the error to <0.05 °C in the temperature range −65 to 100 °C (7). Linearity is also improved using a constant collector current, as pointed out previously. An example of an actual circuit is shown in Fig. 2. In this circuit, the operational amplifier drives the base-emitter voltage to maintain a constant collector current. By applying a square-wave current and measuring the amplitude of the resulting square-wave base-emitter voltage, a linear output proportional to the absolute temperature is obtained, as expected from Eq. 7(12). Further improvement in accuracy can be attained by employing a curve fitting with three-point calibration, the error due to the nonlinearity can be reduced to 0.01 °C in the temperature range of −50 to 125 °C (14).

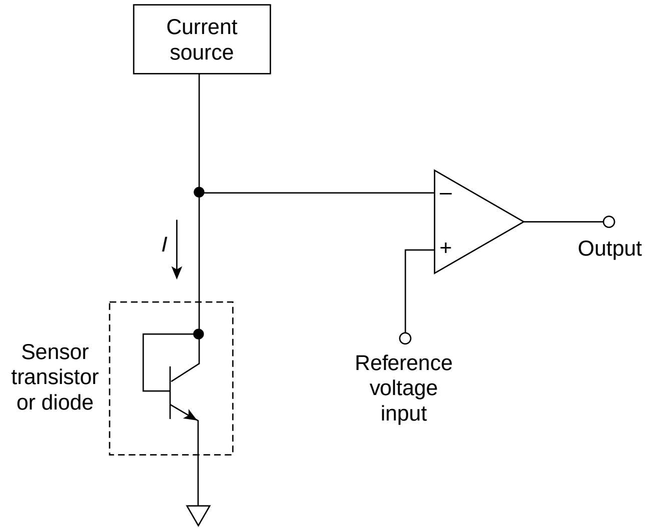


Figure 1. A fundamental interfacing circuit of a thermometer making use of a transistor or a diode as a temperature sensor to provide a voltage output proportional to temperature, with a zero voltage output at a specific temperature dependent on the reference voltage selected.

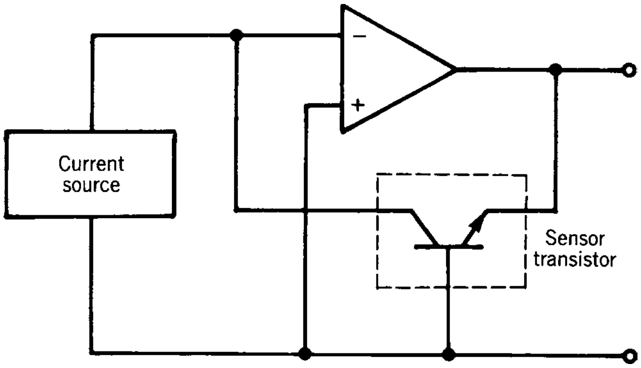


Figure 2. A circuit for constant collector current operation in a sensor transistor.

Three-terminal monolithic IC temperature sensors that provide a voltage output proportional to temperature using the Celsius scale are commercially available, examples being LM45 (National Semiconductor) and AD22100/22103 (Analog Devices). The LM45 device operates using a single power supply voltage in the range 4–10 V, and provides a voltage output that corresponds to the temperature in degrees Celsius multiplied by a factor of 10 mV, for example, 250 mV = 25 °C. The AD22100 and AD22103 devices provide a ratiometric output, that is, the output voltage is proportional to the temperature multiplied by the power supply voltage. For example, AD22100 has a sensitivity of 22.5 mV/ °C giving output voltages of 0.25 V at −50 °C and 4.75 V at 150 °C when the power supply voltage is 5.0 V.

Two matched transistors operated using different collector currents can be used to obtain an output proportional to the absolute temperature (15). The difference in the base-emitter voltages of the two transistors is a linear function of temperature, as shown in Eq. 7. Convenient two-terminal current-output devices using this technique are commercially available. Figure 3 shows an idealized scheme representing such devices. If the transistors Q_1 and Q_2 are assumed to be identical and have a high common-emitter current gain, their collector currents will be equal, and will constrain the collector currents Q_3 and Q_4. If Q_3 has r-fold base-emitter junctions, and each one is identical

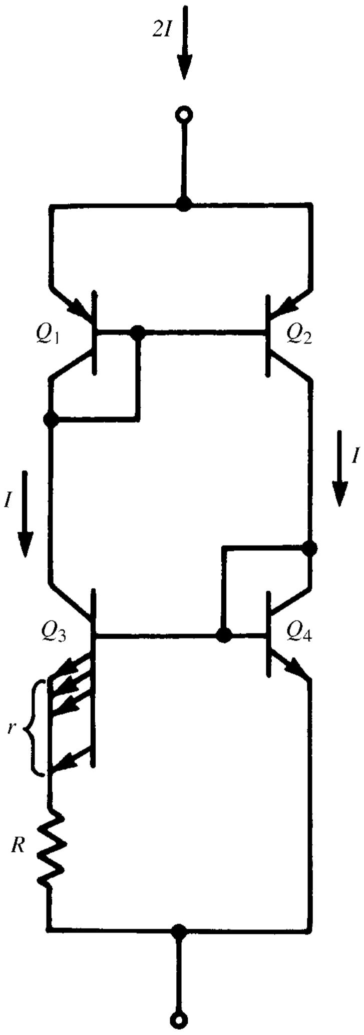


Figure 3. An idealized scheme of a two-terminal IC temperature sensor that provides a current output proportional to the absolute temperature.

to that of Q_4, the emitter current of a junction in Q_3 is $1/r$ that of Q_4. From Eq. 7, the voltage across resistance R is obtained from

$$RI = (kT/q)\ln r \tag{8}$$

Thus, the total current, $2I$, is proportional to the absolute temperature. Although the actual components are not ideal, practical devices are available as monolithic ICs, such as AD590 and AD592 (Analog Devices) (16). In these devices, $r = 8$ and R is trimmed to have a sensitivity of about 1 A/K. The output current is unchanged in the supply-voltage range 4.0 to 30 V. A voltage output proportional to the absolute temperature can be obtained by connecting a resistor in series with the ICs. For example, a sensitivity of 1 mV/K is obtained by connecting 1 kΩ resistor in series. By trimming the series resistor, the error in temperature reading can be adjusted to zero at any desired temperature. After trimming, the maximum error depends on the range in temperature under consideration. For example, a maximum error of <0.1, 0.2, and 0.3 °C is obtained for temperature ranges of 10, 25, and 50 °C, respectively (17).

Monolithic temperature sensors that provide a digital output are also commercially available. For example, TMP06 (Analog Devices) sensors provide a pulse-width modulated output. The output voltage assumes either a high or low level, so that the high period ($T1$) remains constant at 40 ms for all temperatures, while the low period ($T2$) varies with temperature. In the normal operation mode, the temperature on the Celsius *scale*, T, is given by

$$T = 406 - [731 \times (T1/T2)] \tag{9}$$

According to Analog Devices' TMP06 data sheet, for an operating supply voltage between 2.7 and 5.5 V, the absolute temperature accuracy is ±1 °C in the temperature range 0–70 °C, with a temperature resolution of 0.02 °C.

The National Semiconductor LM75 device is also a monolithic temperature sensor that provides a digital output. It includes a nine-bit analog-to-digital converter, and provides a serial output in binary format so that the least significant bit corresponds to a temperature difference of 0.5 °C.

Newer devices will come along in the future that may be more appropriate than the ones mentioned here. Information about such devices, together with their data sheets, will be available from the internet sites of manufactures.

APPLICATIONS

Although thermistors are still widely used for thermometry in the medical field, IC temperature sensors have potential advantages over thermistors. Integrated circuit sensors can be fabricated using IC technology encompassing interfacing electronics on a single IC chip, and many general purpose IC temperature sensors are now commercially available.

Current-output-type IC temperature sensors, such as AD590, are convenient for use as thermometer probes for body core and skin temperature measurements. Figure 4 shows a scheme for such a simple thermometer. According to the manufacturer's data sheet, although the sensitivity and zero offset are adjustable independently in this circuit, an accuracy of 0.1 °C is attainable with L- or M-grade AD590 devices using a single-trim calibration if the temperature span is 10 °C or less. If a regulating resistor is included in the probe, interchangeability can be realized. Because of the current output capacity, the resistance of

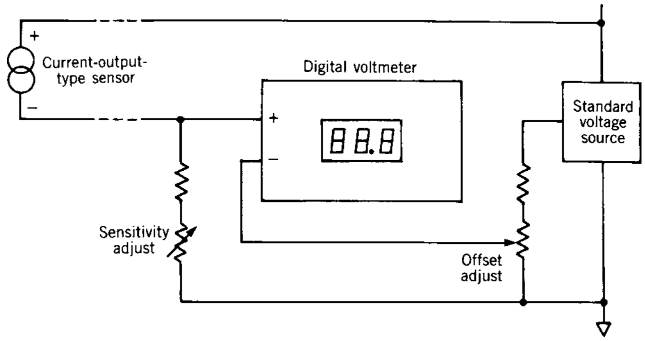


Figure 4. A simple thermometer that makes use of a two-terminal current output-type IC temperature sensor.

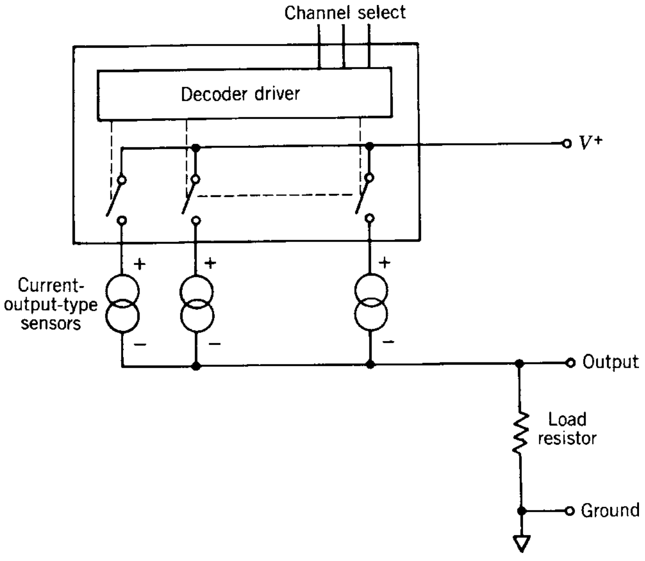


Figure 5. A multiplexing scheme for a current output-type IC temperature sensor.

the cable or connector does not affect the temperature measurement.

This type of device is also convenient for temperature measurements at many other points, especially when the output data are processed using a PC. All the sensors can be connected to a single resistor, as shown in Fig. 5, and by switching the excitation the outputs from each sensor can be multiplexed. To calibrate each sensor individually, all the sensors are maintained at an appropriate temperature, together with a standard thermometer. The outputs from each sensor as well as that from a standard thermometer are input into a PC. Then, the temperature offsets for each sensor can be stored, and all the measurement data can be corrected using these correction factors. Two-point calibration is also realized by using data at two known temperatures. A matrix arrangement of the sensors can be formed using two decoder drivers.

Temperature measurements at many different points can be performed easier using IC temperature sensors that generate serial digital outputs, such as TMP05/TMP06. Connecting these devices as shown in Fig. 6 allows for the realization of a daisy chain operation. When a start pulse is applied to the input of the first sensor, the temperature data from all the sensors is generated serially, so that the temperatures of each sensor are represented in a ratiometric form, which is the ratio of the duration of the high and low output levels for each period. It is a remarkable advantage of this sensor that a thermometer can be realized without using any analogue parts.

An important application of IC temperature sensors is the monitoring of CPU temperatures to protect a system from overheating. The temperature of a CPU chip can be detected by a p–n junction fabricated on the same silicon chip as the CPU, as shown in Fig. 7. The advantage of fabricating the temperature sensor on the CPU chip is to make the temperature measurement accurate enough and to minimize the time delay due to heat conduction so as to prevent overheating. The CPU can be protected from overheating by controlling a cooling fan or by slowing down the clock speed. Interfacing devices for this purpose are commercially available. For example, the MAX6656 (Dallas Semiconductor) device can detect temperatures at three locations, such as the CPU, the battery, and the circuit board, and the output can be used to control a cooling fan. To control the clock frequency, a specially designed frequency generator can be used. For example, the AV9155 (Integrated Circuit Systems) device allows for a gradual transition between frequencies, so that it obeys the CPU's cycle-to-cycle timing specifications.

FUTURE

It is ∼25 years since convenient IC temperature sensors were introduced for scientific and industrial temperature measurements. In medicine, the application of this type of sensor is in its infancy. There are many applications where

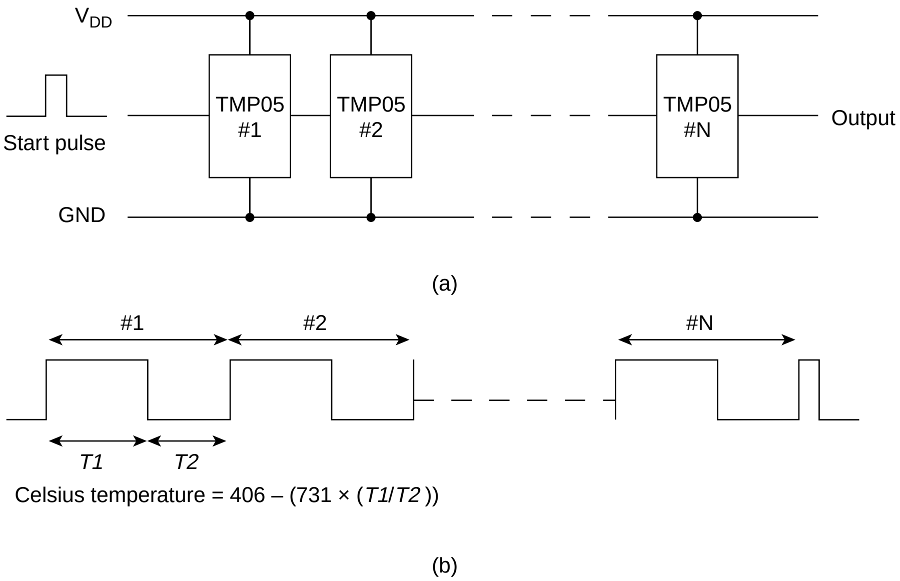


Figure 6. (a) The connecting scheme for a daisy chain operation of a serial-digital-output-type temperature sensor, and (b) the output waveform. The temperature using the Celsius scale at each sensor can be determined from the ratio of the duration of the highest and lowest points in each cycle.

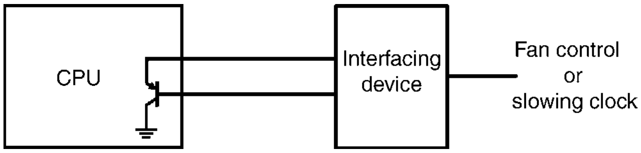


Figure 7. A scheme for monitoring the temperature of a CPU to protect it from overheating by fan control or by slowing down the clock.

these sensors can be used effectively, and undoubtedly their use will be wide spread in the near future.

Digital output IC temperature sensors show the most promise. Using such sensors, thermometers can be made without using analog components, and digital signals are convenient when a photocoupler is used for isolation.

Medical thermometry requires a relatively high degree of accuracy within a narrow temperature range. An absolute accuracy of 0.1 °C is required for body temperature measurements. However, this is hard to attain without individual calibration using most temperature sensors. While adjustment of the trimmer resistor has been used in many thermometer units, correcting data using a PC employing initially obtained correction factors will be much simpler, especially when many sensors are used, and digital output IC temperature sensors are advantageous for such a purpose.

Fabricating different types of sensors, such as force and temperature sensors, in one chip, and then applying them in robot hands to mimic all the sensing modalities of human skin, is another promising field. In such applications, the digital output capability will be a great advantage.

BIBLIOGRAPHY

1. Mackay RS. Endoradiosonde. Nature (London) 1957;179: 1239–1240.
2. Harris H. Concerning a thermometer with solid-state diodes. Sci Amer 1961;204(6):192.
3. MacNamara AG. Semiconductor diodes and transistors as electrical thermometers. Rev Sci Instrum 1963;33: 1091–1093.
4. Cohen BG, Snow WB, Tretola AR. GaAs p-n junction diodes for wide range thermometry. Rev Sci Instrum 1963; 34:1091–1093.
5. Griffiths B, Stow CD, Syms PH. An accurate diode thermometer for use in thermal gradient chambers. J Phys E 1974; 7:710–714.
6. Felimban AA, Sandiford DJ. Transistors as absolute thermometers. J Phys E 1974;7:341–342.
7. Davis CE, Coates PB. Linearization of silicon junction characteristics for temperature measurement. J Phys E 1977;10:613–619.
8. O'Neil P, Derrington C. Transistors—a hot tip for accurate temperature sensing. Electronics 1979;52(21):137–141.
9. Sah C, Noyce RN, Shockley W. Carrier generation and recombination in *p-n* junctions and p-n junction characteristics. Proc IRE 1957;45:1228–1243.
10. Sah C. Effect of surface recombination and channel on p-n junction transistor characteristics. IRE Trans Electron Devices 1962;ED9:94–108.
11. Sclar N, Pollock DB. On diode thermometers. Solid State Electron 1972;15:473–480.
12. Verster TC. p-n junction as an ultralinear calculable thermometer. Electron Lett 1968;4:175–176.
13. Ruhle RA. Solid-state temperature sensor outperforms previous transducers. Electronics 1975;48(6):127–180.
14. Ohte A, Yamagata M. A precision silicon transistor thermometer. IEEE Trans Instrum Meas 1977;IM-26:335–341.
15. Vester TC. Dual transistor as thermometer probe. Rev Sci Instrum 1969;40:174–175.
16. Timko MP. A two-terminal IC temperature transducer. IEEE J Solid-State Circuits 1976;SC-11:784–788.
17. Sheingold DH, editor. Transistor Interfacing Handbook, A Guide to Analog Signal Conditioning. Norwood, MA: Analog Devices; 1980. p 153–177.

Further Reading

Sze SM. Semiconductor Devices—Physics and Technology. New York: John Wiley & Sons; 1985.

Togawa T, Tamura T, Öberg PA. Biomedical Transducers and Instruments. Boca Raton, FL: CRC Press; 1997.

Moore BD. IC temperature sensors find the hot spot. EDN July 2/98, 1998; 99–110.

Frank R. Semiconductor junction thermometers. In: Webster JG, editor. The Measurement, Instrumentation, and Sensors Handbook. Boca Raton, FL: CRC Press; 1999. p 32/74–32/87.

See also CAPACITIVE MICROSENSORS FOR BIOMEDICAL APPLICATIONS; ION-SENSITIVE FIELD EFFECT TRANSISTORS; THERMOMETRY.

INTERFERONS. See IMMUNOTHERAPY.

INTERSTITIAL HYPERTHERMIA. See HYPERTHERMIA, INTERSTITIAL.

INTRAAORTIC BALLOON PUMP

PETER WELLER
City University
London, United Kingdom
DARREN MORROW
Royal Adelaide Hospital
Adelaide, Australia

INTRODUCTION

The heart is a pump made of cardiac muscle or myocardium. It has four pumping chambers, namely, a right and left atrium and a right and left ventricle. The atria act as primer pumps for the ventricles. The right ventricle pumps deoxygenated blood returning from the body through the pulmonary artery and into the lungs. This is called the pulmonary circulation. The left ventricle pumps oxygenated blood returning from the lungs through the aorta and into the rest of the body. This is called the systemic circulation.

The heart also has four one-way valves that prevent the backward flow of blood. The tricuspid valve lies between the right atrium and right ventricle while the pulmonary valve lies between the right ventricle and the pulmonary artery. Similarly, the mitral valve lies between the left atrium and the left ventricle while the aortic valve lies between the left ventricle and the aorta.

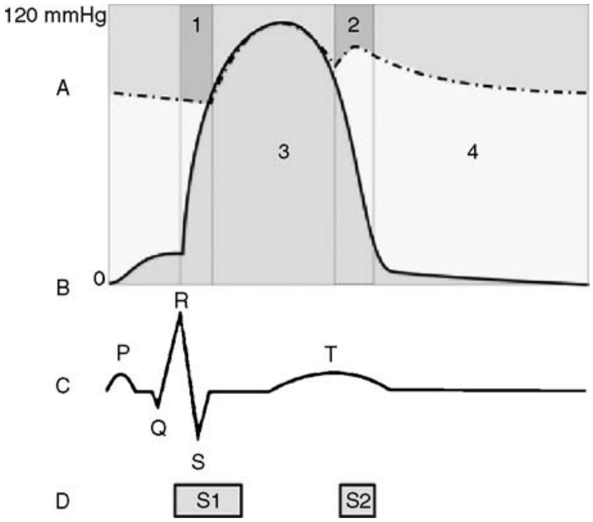


Figure 1. The relationship between the aortic pressure (A – dashed line), the ventricular pressure (B – solid line), the electrocardiogram (C) and the heart sounds (D). Region 1 = isovolumetric contraction and Region 2 = isovolumetric relaxation, Region 3 (green) = tension time index (TTI) and Region 4 (yellow) = diastolic pressure time index (DPTI). S1 represents the closing of mitral and tricuspid valves, S2 represents the closure of aortic and pulmonary valves.

CARDIAC CYCLE

The heart pumps rhythmically. The cardiac cycle is the sequence of events that take place in the heart during one heartbeat (Fig. 1). Thus, the duration of the cardiac cycle varies inversely with the heart rate. At a typical resting heart rate of 60 beats per minute (bpm), the cardiac cycle lasts 1 s or 1000 ms.

Mechanical Events

One cardiac cycle consists of a period of contraction called systole followed by a period of relaxation called diastole. The duration of systole, called the systolic time interval (STI), is relatively constant, but the duration of diastole, called the diastolic time interval (DTI), varies with the heart rate. Thus, when the heart rate increases, the DTI shortens.

When the left ventricle contracts, the pressure in the left ventricle rises above the pressure in the left atrium and the mitral valve closes. Soon afterward, the pressure in the left ventricle rises above the pressure in the aorta and the aortic valve opens. Blood flows from the left ventricle into the aorta. The period between closing of the mitral valve and opening of the aortic valve is called isovolumetric contraction.

When the left ventricle relaxes, the pressure in the left ventricle falls below the pressure in the aorta and the aortic valve closes. This causes a momentary drop in pressure in the aorta called the dichrotic notch. The period between the opening and closing of the aortic valve is called ventricular ejection. Soon afterward, the pressure in the left ventricle falls below the pressure in the left atrium and the mitral valve opens. Blood flows from the left atrium into the left ventricle. The period between the closure of the aortic valve and opening of the mitral valve is called isovolumetric relaxation.

The left atrium contracts and relaxes just before the left ventricle. This boosts the blood flow from the left atrium into the left ventricle.

These events are mirrored in the right ventricle and right atrium. However, the pressures in the pulmonary circulation are much lower than the pressures in the systemic circulation.

The movements of the chambers, valves and blood can be imaged noninvasively using ultrasound and this is called an echocardiogram.

Electrical Events

The rhythmical pumping of the heart is caused by waves of electrical impulses that spread through the myocardium from the atria to the ventricles. A recording of these waves is called an electrocardiogram (ECG). The P wave represents atrial contraction. The R wave represents ventricular contraction and signals the beginning of systole. The T wave represents ventricular relaxation and signals the beginning of diastole.

Acoustic Events

The opening and closing of the valves in the heart creates sounds that can be heard at the surface of the chest using a stethoscope. A recording of these sounds is called a phonocardiogram. The first heart sound (S1) represents closure of the mitral and tricuspid valves and signals the beginning of systole. The second heart sound (S2) represents closure of the aortic and pulmonary valves and signals the beginning of diastole.

MYOCARDIAL OXYGEN BALANCE

The systemic circulation delivers oxygenated blood to the body. Body tissues use oxygen to generate energy from the oxidation of fuels. All tissues, including the myocardium, need energy to function. The net delivery of oxygen to the myocardium is called the myocardial oxygen balance.

$$M_{\mathrm{OB}} = M_{\mathrm{OS}} - M_{\mathrm{OD}}$$

where M_{OB} = myocardial oxygen balance, M_{OS} = myocardial oxygen supply, M_{OD} = myocardial oxygen demand. In the healthy heart the myocardial oxygen balance is positive, that is supply exceeds demand. In the failing heart the balance can be negative, that is demand exceeds supply.

Myocardial Oxygen Supply

The main blood supply of the myocardium comes from the two coronary arteries and their branches. The small amount of blood that reaches the myocardium transmurally from within the chambers of the heart is insignificant. The coronary arteries arise from the aorta just beyond the aortic valve and ramify within the myocardium. Myocardial oxygen supply depends on the coronary blood flow and the amount of oxygen that can be extracted from the blood.

When the heart contracts the coronary arteries are compressed and the coronary blood flow is decreased. The net driving force for coronary blood flow is called the coronary perfusion pressure.

$$C_{\mathrm{PP}} = A_{\mathrm{P}} - V_{\mathrm{P}}$$

where C_{PP} = coronary perfusion pressure, A_{P} = aortic pressure, V_{P} = ventricular pressure. The coronary circulation is unique because more blood flows during diastole when the ventricular pressure is low than during systole when the ventricular pressure is high. Thus, the coronary blood flow depends on the coronary perfusion pressure, the diastolic time interval and the patency of the coronary arteries. Myocardial oxygen supply is represented by the area between the aortic pressure wave and the left ventricular pressure wave, called the diastolic pressure time index (DPTI).

Myocardial Oxygen Demand

The myocardium uses energy to perform the work of pumping. The work performed by the heart can be estimated by the mean aortic blood pressure multiplied by the cardiac output. Myocardial oxygen demand depends on the heart rate, the systolic wall tension and the cardiac contractility. Systolic wall tension is developed during isovolumetric contraction and depends upon the preload, the afterload and the wall thickness. The preload is the degree to which the left ventricle is filled before it contracts, that is the left ventricular end diastolic volume. The afterload is the pressure in the aorta or the systemic vascular resistance against which the left ventricle contracts. Myocardial oxygen demand is represented by the area under the left ventricular pressure curve, called the tension time index (TTI).

The myocardial oxygen balance is represented by the ratio DPTI:TTI.

THE PATHOPHYSIOLOGY OF LEFT VENTRICULAR PUMP FAILURE

When the left ventricle begins to fail as a pump, the cardiac output falls. Compensatory physiological mechanisms bring about an increase in left ventricular end diastolic volume, heart rate, and systemic vascular resistance. The result is an increase in preload and afterload with a decrease in coronary blood flow. Thus, the myocardial oxygen demand increases while the myocardial oxygen supply decreases. There may come a point when demand exceeds supply resulting in a negative myocardial oxygen balance. The left ventricle is then deprived of oxygen and cannot generate sufficient energy to do the work required of it. The pump failure is therefore exacerbated and this can precipitate a downward spiral of decline eventually ending in death. The therapeutic goal is to reverse this decline and help the failing left ventricle to recover by restoring a positive myocardial oxygen balance. Diuretics to decrease the preload, inotropic drugs to increase the myocardial contractility and vasodilators to decrease the preload and afterload are the mainstay of treatment. However, in the most severely ill patients, pharmacological measures alone may be insufficient and it is in these extreme circumstances that counterpulsation therapy may be effective.

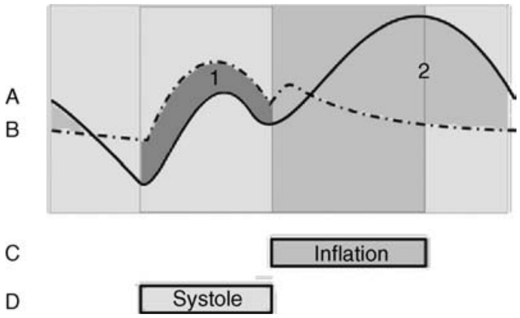


Figure 2. The relationship between the ventricular pressure with counterpulsation (A – solid line), without counterpulsation (B-dashed line), IABP balloon inflation (C) and ventricular systole (D). Region 1 (green) = systolic unloading and Region 2 (yellow) = diastolic augmentation.

THE PRINCIPLE OF COUNTERPULSATION

The principle of counterpulsation is the incorporation of an additional pump into the systemic circulation in series with the left ventricle. The pump is operated in synchrony, but out of phase, with the cardiac cycle. Pump systole occurs during ventricular diastole and pump diastole occurs during ventricular systole.

The primary physiological effects of counterpulsation are twofold (Fig. 2): A decrease in the aortic pressure during systole (called systolic unloading). This is evidenced by a decrease in the end diastolic pressure (EDP), the peak systolic pressure (PSP) and the mean systolic pressure (MSP). An increase in the aortic pressure during diastole (called diastolic augmentation). This is evidenced by an increase in the mean diastolic pressure (MDP).

Systolic unloading reduces the work of the left ventricle because it pumps against a lower pressure. This decreases myocardial oxygen demand. Diastolic augmentation increases coronary blood flow because it increases the coronary perfusion pressure. This increases myocardial oxygen supply. Thus, the myocardial oxygen balance is improved.

Among the secondary physiological effects of counterpulsation are increases in the stroke volume (SV, the volume of blood pumped with each heartbeat), the CO (equal to the SV multiplied by the heart rate) and the blood flow to the other vital organs.

HISTORICAL PERSPECTIVE

Counterpulsation was first described in theory in 1958 by Harken (1). It was to be achieved by cannulating the femoral arteries, rapidly withdrawing a set volume of blood during systole and rapidly reinfusing the same volume of blood during diastole. Clauss et al. (2) reported this in clinical practice in 1961 but it was unsuccessful because the rapid movements of blood were difficult to implement and caused severe hemolytic damage to the red blood cells.

In 1958, Kantrowitz and McKennin (3) described counterpulsation achieved by wrapping a part of the diaphragm around the thoracic aorta and stimulating the phrenic nerve, causing contraction of the diaphragm, during diastole. Moulopoulos et al. (4) and Clauss et al.(5) described counterpulsation achieved by intra-aortic balloon pumping in 1962. Operative insertion of the balloon through a surgically exposed femoral artery was necessary. It was inflated during diastole and deflated during systole. In 1968, Kantrowitz et al. (6) reported a successful clinical study.

In 1963, Dennis et al.(7) described external counterpulsation achieved using a pneumatic compression garment that enclosed the legs and lower torso. It was inflated during diastole and deflated during systole. It was reported to be as successful as the IABP in clinical studies but it is not commonly used. Kantrowitz et al. (8) described counterpulsation achieved by a permanently implantable intra-aortic balloon in 1972. It was unsuccessful in clinical practice because there remained the need for a connection to an external pump and this provided a portal of entry for infection.

In 1979, following the development of thinner catheters, percutaneous insertion of the intra-aortic balloon through a femoral artery puncture was introduced. This could be performed at the bedside and avoided the need for a surgical operation in most patients. Consequently, intra-aortic balloon pumping became the most widely adopted method of counterpulsation.

CLINICAL APPLICATIONS

Indications

The IABP was first used clinically in 1968 by Kantrowitz to support patients with cardiogenic shock after acute myocardial infarction(9). During the 1970s the indications broadened (Table 1) and by 1990 ~70,000 pump procedures were performed worldwide each year (10) although there is wide variation between different countries and centres. The IABP support has been used successfully in patients with left ventricular failure or cardiogenic shock from many causes including myodarditis, cardiomyopathy, severe cardiac contusions and drug toxicity but the commonest are myocardial infarction and following cardiac surgery. The trend has been a move away from hemodynamic support in pump failure towards the treatment, and even prophylaxis of, acute myocardial ischaemia. Patients can be maintained on the IABP for hours, days or even weeks, particularly when used as a bridge to cardiac transplantation or other definitive treatment (11). Of those who survive to hospital discharge, long-term survival is satisfactory (12).

An early series of 747 IABP procedures in 728 patients between 1968 and 1976 was reported by McEnany et al. (13). Over the course of the study, they observed that cardiogenic shock or chronic ischaemic left ventricular failure as the indication for IABP fell from 79 to 26% of patients whilst overall in-hospital survival rose from 24 to 65% of patients. They also noted an increase from 38 to 58% of patients undergoing cardiac surgery following IABP insertion. They postulated that broadened indications for, and earlier insertion of, the IABP together with more aggressive surgical treatment of any underlying cardiac lesion led to the improvement in survival. In the later Benchmark Registry of nearly 17,000 IABP procedures performed in 203 hospitals worldwide between 1996 and 2000, the main indications were support for coronary angioplasty (21%), cardiogenic shock (19%), weaning from cardiopulmonary bypass (16%), preoperative support in high risk patients (13%), and refractory unstable angina (12%) (14). The overall in-hospital mortality was 21%.

High risk patients undergoing cardiac surgery may have a better outcome if treated preoperatively with IABP therapy. In a series of 163 patients with a left ventricular ejection fraction of <0.25 and undergoing coronary artery bypass grafting (CABG), the 30 day mortality was reduced from 12 to 3% (15). Similar results were obtained in a small randomized study (16). In a series of 133 patients who underwent CABG off cardiopulmonary bypass between 2000 and 2003, the use of adjuvant preoperative IABP therapy in the 32 highest risk patients led to outcomes comparable with the lower risk patients (17). The use of IABP therapy to improve outcome after coronary angioplasty for acute myocardial infarction remains controversial. Early studies suggested an improved outcome (18–20), but two recent large randomized trials have shown no benefit in haemodynamically stable patients (21,22). A report from the SHOCK Trial Registry showed that the in-hospital mortality in patients with cardiogenic shock after acute myocardial infarction could be reduced from 77% to 47% by combined treatment with thrombolysis and IABP, particularly when followed by coronary revascularization (23).

IABP therapy is used infrequently in children, who commonly suffer from predominantly right ventricular failure associated with congenital heart disease. The greater elasticity of the aorta may limit diastolic augmentation and the more rapid heart rate may make ECG triggering difficult. Echocardiographic triggering has been used as an effective alternative (24,25). Survival rates of 57% (26) and 62% (27) have been reported in small series of carefully selected patients.

Contraindications

The only absolute contraindications to IABP therapy are severe aortic regurgitation and aortic aneurysm or dissection. In patients with severe aorto-iliac vascular disease

Table 1. Indications for IABP Therapy

Left ventricular failure or cardiogenic shock	Preoperative support before cardiac or non-cardiac surgery
Refractory unstable angina or ischaemic ventricular arrhythmias	Adjunct to coronary angioplasty or thrombolyis
Weaning from cardiopulmonary bypass	Adjunct to off-bypass cardiac surgery
Bridge to cardiac transplantation	

Table 2. Complications of IABP Therapy

Vascular	Balloon-Related	Other
Hemorrhage	Gas embolism	Infection
Aortoiliac dissection or perforation	Entrapment	Thrombocytopenia
		Paraplegia
Limb ischaemia		
Visceral ischaemia		

the balloon should not be inserted through the femoral artery.

Complications

The IABP therapy continues to cause a significant number of complications (Table 2), but serious complications are uncommon and directly attributable deaths are rare. Nevertheless, some have argued against its indiscriminate use, feeling that for many patients the risks outweigh the benefits. Kantrowitz reported rates of 41 and 4% for minor and major complications, respectively, in his series of 733 patients. Of these, 29% were vascular (including 7% hemorrhagic) and 22% were infections (28). Vascular complications include haemorrhage from the insertion site and lower limb ischaemia caused by the balloon catheter or sheath occluding the iliac or femoral artery. The vascular status of the lower limbs should be observed closely in patients on IABP therapy. Ischaemia may resolve when the catheter or sheath is removed but surgical intervention including femoral thromboembolectomy, femorofemoral bypass or even amputation is required in up to half of cases (29).

Several risk factors for vascular complications have been identified. They are female gender, diabetes, hypertension, peripheral vascular disease, obesity, old age, sheathed insertion, percutaneous insertion, and insertion via the femoral artery compared to directly into the ascending aorta (13,14,19,28–30). In one study of patients with peripheral vascular disease, the rate of vascular complications was 39% for percutaneous insertion compared to 18% for open insertion (29).

Complications caused by perforation of the balloon are rare, but potentially serious. Embolization of the helium shuttle gas can result in stroke or death. Coagulation of blood within the balloon can result in balloon entrapment. In this situation the instillation of thrombolytic agents may allow the balloon to be retrieved percutaneously but otherwise open surgery is required. It is therefore mandatory to remove the balloon immediately if any blood is detected within the pneumatic system.

Thrombocytopenia (a reduction in the number of platelets) developed in one-half of patients but they rapidly recovered when the balloon was removed (31).

EQUIPMENT FOR CLINICAL APPLICATION

The IABP consists of a balloon catheter and movable drive console. A monitor on the drive console displays the arterial pressure wave and the ECG. Commercial consoles have controls that allow the operator to select the assist ratio and trigger mode and adjust the timing of inflation and deflation and the inflation volume of the balloon. The drive console also contains a helium tank for balloon inflation and a battery as a backup power source in the event that the mains electricity supply is interrupted. In common with medical equipment the IABP console conforms to international safety standards. Figure 3 shows a current commercial model.

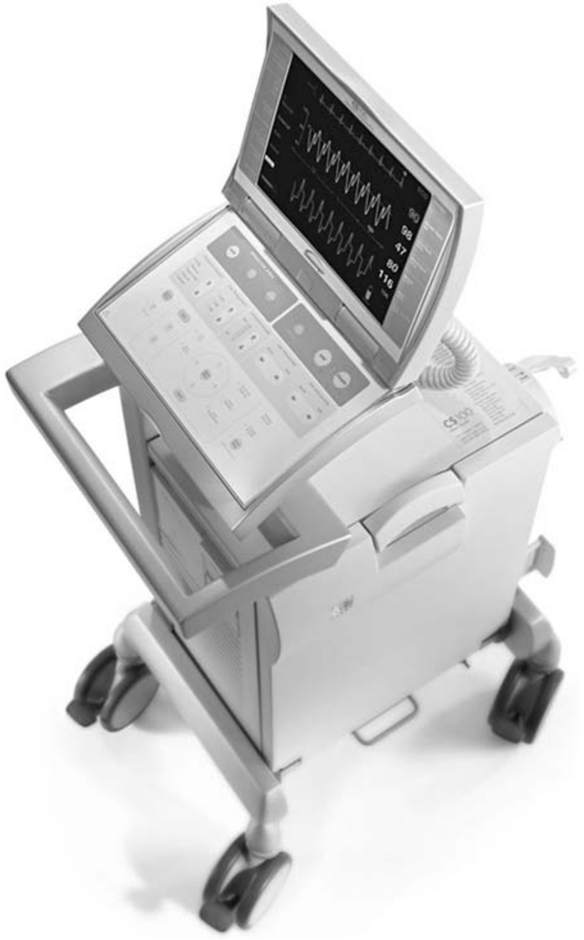


Figure 3. A commercial IABP device. (Courtesy of Datascope Corp.)

The balloon is made of inelastic polyurethane and is cylindrical in shape. Balloons are available in volumes from 25 to 40 cm^3 and the correct size is selected according to the height of the patient. The balloon is mounted at the end of a double-lumen catheter. Modern catheters have an outer diameter of 7–8 French gauge. The inner lumen is open at the tip to allow insertion over a guidewire and direct measurement of the aortic blood pressure after the guidewire is removed. The outer lumen forms a closed system connecting the balloon to a pneumatic pump chamber within the drive console. Two views of a current balloon catheter are shown in Fig. 4.

The balloon catheter is most commonly inserted percutaneously through the femoral artery in the groin over a guidewire using a traditional or modified Seldinger technique. An intra-arterial sheath is used to secure and

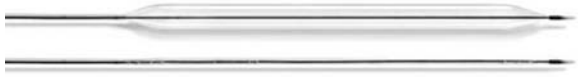


Figure 4. An IABP catheter. (Courtesy of Datascope Corp.) In both inflated (top image) and uninflated (bottom image) modes.

protect the access site before insertion of the balloon catheter. However, because the sheath has a larger outer diameter than the balloon catheter it can increase the risk of vascular complications and sheathless insertion has now been introduced. Under fluoroscopic guidance the balloon is positioned in the descending thoracic aorta just beyond the origin of the left subclavian artery. Alternatively, the balloon can be inserted through the iliac, axillary or subclavian arteries or directly into the ascending aorta during open surgery.

A shuttle gas is pumped back and forth between the pump chamber and the balloon to cause inflation and deflation. The ideal shuttle gas would have a low density combined with a high solubility in blood. A dense gas is slow to move along the catheter. This introduces a significant delay between the opening of the valves in the pump chamber and the inflation or deflation of the balloon making correct timing more difficult to achieve. An insoluble gas is unsafe if the balloon was to leak or burst allowing it to escape into the blood. There it may form bubbles leading to potentially fatal gas embolism. Originally, carbon dioxide was used as the shuttle gas. It is a dense gas but dissolves easily in blood. More recently, helium has been used as the shuttle gas. It is a less dense gas but dissolves less easily in blood.

Pumping is initiated with an assist ratio of 1:2, which means that one in every two heartbeats is assisted. This allows the arterial pressure wave of each assisted beat to be compared with an unassisted beat to facilitate correct timing.

Pumping is continued with an assist ratio of 1:1, which means that every beat is assisted. As the patient recovers, they can be weaned from the IABP by periodically decreasing the assist ratio until pumping is eventually discontinued.

Weaning can also be achieved by periodically decreasing the inflation volume of the balloon. However, this can lead to problems with blood clotting in the folds of the underinflated balloon and it is not commonly used.

CONTROL OF IABP

The inflation/deflation cycle of the intra-aortic balloon pump is controlled by a closed loop circuit as illustrated in Fig. 5.

The physiological variables required for determining triggering are measured, filtered and converted into digital signals. These are used in enable the control strategy to determine the appropriate inflation and deflation times of the balloon. This information is then passed to the pneumatic circuit for balloon operation. The results of this strategy are then feed back to the console via a new set of patient variables. The actions of the controller are displayed on the console monitor. Additionally the control strategy can be set by the clinician.

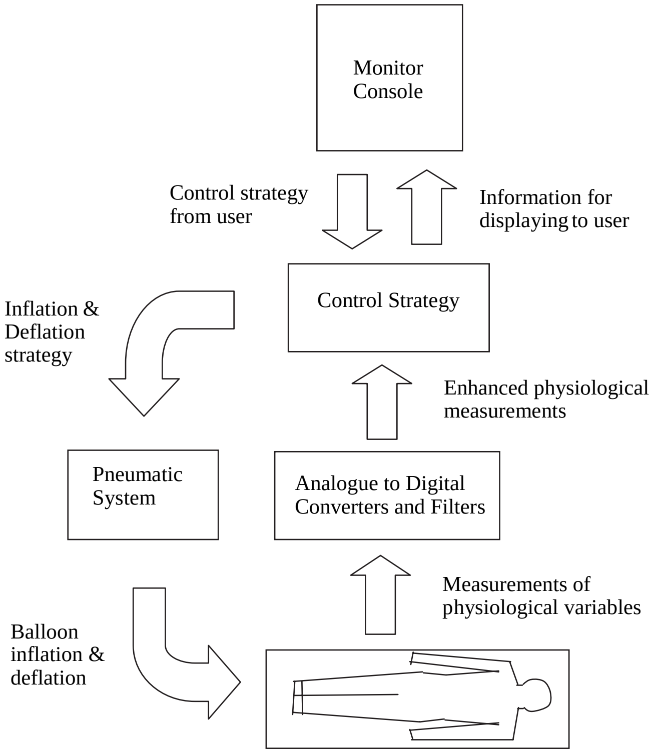

Monitor Console
Control strategy from user
Information for displaying to user
Control Strategy
Inflation & Deflation strategy
Enhanced physiological measurements
Pneumatic System
Analogue to Digital Converters and Filters
Measurements of physiological variables
Balloon inflation & deflation

Figure 5. Closed-loop circuit for IABP control. The arrows indicate the flow of information for controlling the IABP.

TRIGGERING

The drive console requires a trigger signal to synchronise with the cardiac cycle. The trigger signal indicates the start of ventricular systole.

Commercial consoles have several modes of triggering: The ECG trigger, where the trigger signal is the R wave of the ECG. This is the most commonly used mode of triggering. Blood pressure trigger, where the trigger signal is the upstroke of the arterial pressure wave. This is used when the ECG signal is too noisy to allow reliable R wave recognition. Pacer trigger, where the trigger signal is the pacing spike of the ECG. This is used when the patient has an implanted cardiac pacemaker. Internal trigger, where the trigger signal is generated internally by the console at a set rate. This is used during cardiac surgery when the heartbeat is temporarily arrested.

TIMING

The safety and efficacy of balloon pumping is dependent upon the correct timing of balloon inflation and deflation during the cardiac cycle. Inflation should occur during the isovolumetric relaxation period of ventricular diastole. Too early inflation is unsafe because it overlaps the ejection period of ventricular systole, impedes ejection and increases the work of the heart. Too late inflation is less effective because it reduces diastolic augmentation.

Safe and effective inflation timing is fairly easy to achieve because the systolic time interval is relatively constant, regardless of heart rate or rhythm. The operator is trained to adjust the time of inflation until it visibly corresponds with the dichrotic notch on the arterial pressure wave display.

Deflation should occur at the end of ventricular diastole or during the isovolumetric contraction period of ventricular systole. Too late deflation is unsafe because it overlaps the ejection period of ventricular systole, impedes ejection and increases the work of the heart. Too early deflation is less effective because it reduces diastolic augmentation.

Safe and effective deflation timing is more difficult to achieve because the length of diastole is variable, depending on the heart rate and rhythm. The operator is trained to adjust the time of deflation until it visibly reduces the EDP and PSP on the arterial pressure wave display. Commercial consoles have two modes of timing:

Conventional Timing

Under conventional timing, the balloon is inflated and deflated at set time intervals after the R wave is detected (called manual timing). This copes badly with alterations in heart rate of >10 bpm. The problem is that when the heart rate increases, the diastolic time interval shortens and the balloon now deflates too late. Conversely, when the heart rate decreases, the diastolic time interval lengthens and the balloon deflates too early.

Conventional timing can be improved by predicting the duration of the current cardiac cycle by averaging the R–R interval of the previous 5–20 heartbeats. The balloon is then deflated at a set time interval before the next R wave is predicted to arrive (called predictive timing). This copes fairly well with alterations in heart rate and is the most commonly used mode of timing.

Both manual and predictive timing cope badly with alterations in heart rhythm, when the length of diastole can vary from beat to beat in an entirely unpredictable way. Unfortunately, cardiac arrhythmias such as atrial fibrillation and frequent ectopic beats are common in these patients making optimal timing difficult to achieve.

Real Timing

Under real timing (also called R wave deflation), the balloon is deflated when the R wave is detected and inflated a set time interval later. This copes very well with alterations in heart rate and rhythm but it tends to cause too late deflation. The problem is that when the R wave is detected there may be insufficient time to fully deflate the balloon before overlapping the ejection period.

The R wave deflation is usually used as a safety mechanism in conjunction with conventional timing. It ensures that the balloon is deflated if an R wave arrives unexpectedly due to a sudden change in heart rate or rhythm.

OPTIMIZATION

As was seen above the standard timing strategies both suffer from problems in certain scenarios and thus control of IABPs are not efficient in providing the best patient treatment. A natural solution to this is to develop timing strategies that optimize the balloon inflation regime according to the current patient condition. Several teams have reported work in this area.

Jaron et al. in 1979 (32) developed a multielement mathematical model of the canine circulation and the IABP. They expressed inflation and deflation times in terms of the total duration of inflation (DUR) and the time from the R wave to the middle of pump systole (TMPS). Duration of inflation and TMPS were expressed in terms of percentages of the duration of the cardiac cycle.

The model was validated by comparison with anesthetized dogs. They varied DUR and TMPS and measured the effects on EDP, MDP, and CO. Each of the three dependent variables (z axis) were plotted against the two independent variables (x and y axes) to create a three-dimensional (3D) urface. In general, there was considerable similarity between the surfaces obtained from the model and from the dogs. In particular, the similarity was closer for measurements of pressure (EDP, MDP) than for measurements of flow (CO).

Their results indicated that the locations of the desired optima for EDP (75% DUR, 45% TMPS) & CO (55% DUR, 75% TMPS) did not coincide. Furthermore, some combinations of DUR and TMPS within the range used clinically produced detrimental effects. Because not all variables could be optimised at same time, they suggested that the choice of timing settings involved balancing the clinical needs of the patient.

Jaron et al. in 1983 (33) subsequently developed a lumped model of the canine circulation and the IABP. It was validated by comparison with their previous model. They varied the timing of inflation and deflation, the speed of inflation and deflation and the volume of the balloon. The speed was either fast or slow, taking 7% or 33% of the duration of the cardiac cycle, respectively. The fast speed represented the ideal console with near instantaneous balloon inflation and deflation. The slow speed represented commercial consoles with significant inflation and deflation delay due to the movement of the shuttle gas. They measured the effects on EDP, SV, and coronary blood flow.

Inflation at end systole maximized SV and coronary blood flow for fast and slow speeds. Deflation at end diastole minimized EDP for fast speeds. At slow speed, deflation timing involved a trade-off between decreased EDP with early deflation and increased SV and coronary blood flow with late deflation. The overall benefit of the IABP was greater with fast speeds than slow speeds. It was also proportional to balloon volume.

In later experiments (34), they classified dependent variables as either internal, reflecting myocardial oxygen demand, or external, reflecting myocardial oxygen supply. Internal variables measured were TTI and EDP. External variable measured were SV and MDP. They showed that early deflation minimizes internal variables while late deflation maximizes external variables.

Niederer and Schilt in 1988(35) used a mechanical mock circulation and a mathematical model to investigate then influence of timing of inflation and deflation, speed of inflation and deflation and balloon volume on the efficacy of the IABP. Timing was again expressed in terms of percentage of cardiac cycle duration. The default settings of the model were fast inflation at 30% time, fast deflation at 90% time and a volume of 40 cm^3. These were found to produce an increase in SV of 25%, a decrease in left ventricular systolic pressure of 10% and an increase in aortic diastolic pressure of 50%.

The time of inflation was varied between 20% and 50%. This had little effect on SV when the speed was fast, but inflation after 30% caused a slight decrease in SV compared to default when the speed was slow. A time of deflation before 80% caused a slight decrease in SV compared to default. Fast inflation and deflation speeds caused opening and closing shock waves that could be harmful were they to occur in humans. Balloon volume was varied between 10 and 50 mL. There was a nonlinear relationship with SV, but 40 mL was adequate for optimal performance in the mock circulation.

Barnea et al. in 1990 (36) developed a sophisticated computer simulation of the normal, failing and IABP-assisted failing canine circulation. It included simple physiological reflexes involved in the regulation of the cardiovascular system. The failing heart was simulated by reducing the contractility of the normal heart. Myocardial oxygen supply and demand were calculated from the model. They were balanced in the normal circulation and imbalanced in the failing circulation. IABP assistance of the failing circulation was shown to restore the balance.

Sakamoto et al. in 1995 (37) investigated the effect of deflation timing on the efficiency of the IABP in anaesthetized dogs. They compared a deflation time before the R wave (during late diastole) with deflation times after the R wave (during isovolumetric contraction). Deflation during the middle of isovolumetric contraction was the most effective in obtaining optimal systolic unloading.

Morrow and Weller (38) successfully used genetic algorithms and the fitness function proposed by Kane et al.(39) to evolve a fuzzy controller that optimized cardiac assistance in a computer simulation of the IABP-assisted failing heart. The inputs were MDP and PSP and the output was deflation time.

Automatic Control

Kane et al. in 1971 (39) proposed a performance index or fitness function that reflected the overall benefit of IABP assistance at different timing combinations. Inflation and deflation times were expressed in terms of the delay before inflation after the R wave and the duration of inflation. The function included weighted MDP, MSP, and EDP.

$$\text{Fitness} = k_1\text{MDP} + k_2\text{MSP} + k_4\delta(k_3 - \text{EDP})^2$$

$$\text{where,}\quad k_1 = \frac{100}{\text{MDP}_0},\quad k_2 = \frac{100}{\text{MDP}_0},\quad k_3 = \text{EDP}_0,$$

$$k_4 = -500\left(\frac{1}{k_3}\right)^2,\quad \delta = \begin{cases} 1 & \text{if}(k_3 - \text{EDP}) < 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{MDP}_0 = \text{unassisted MDP},\ \text{EDP}_0 = \text{unassisted EDP}$$

It was tested in a mechanical mock circulation and fitness was found to be a unimodal function of delay and duration. An automatic controller was developed that used a gradient descent search algorithm to improve the fitness by adjusting the delay and duration at each heartbeat. The performance of the controller was compared with the performance of R wave deflation in simulated cases of heart failure. The controller was considerably better in moderate heart failure and marginally better in severe failure. In severe failure the trade-off between systolic unloading or diastolic augmentation was particularly apparent. In moderate failure, the fitness function emphasized systolic unloading (early deflation) while in severe failure it adapted to emphasize diastolic augmentation (late deflation). Thus, in severe failure the controller tended to simulate R wave deflation.

Martin and Jaron in 1978 (40) developed a manual controller for the IABP that allowed DUR and TMPS to be adjusted. It was tested successfully on anesthetized dogs and was capable of linking to a computer for automatic control.

Jaron et al. in 1985 (34) suggested that SV and TTI index were suitable choices for a fitness function for the fine adjustment of inflation time. Both MDP and EDP were suitable choices for the adjustment of deflation time. However, later work showed that PSP correlated better with myocardial oxygen demand than EDP.

Barnea (41,42), Smith (43) and their co-workers in 1989 proposed a fitness function for optimal control of deflation time. It included weighted MDP and PSP. The permitted interval for deflation time was −200 ms to +100 ms relative to the predicted arrival of the next R wave. An automatic controller was developed that used a search and approximation algorithm to converge upon the optimum fitness after a number of heartbeats. It was tested successfully on computer simulations and anaesthetised dogs. It was able to follow a moving optimum, both within the same patient over time and between different patients.

Zelano et al. in 1990 (44) developed an automatic controller for the IABP that used different trigger signals. Balloon inflation occurred either upon detection of S2 or at a set time prior to the predicted time of the next S2. Balloon deflation occurred either during the P–R interval at a set time after the P wave or after the R wave. The advantage of using the P wave, R wave and S2 for triggering is that all can be detected in real time. This allows the controller to follow changes in heart rate and rhythm. It was tested successfully in a semiautomatic open loop operation in anaesthetised dogs with a coronary artery tied to simulate a myocardial infarction. They proposed a fitness function for automatic closed loop control. It used weighted MSP and MDP.

Kantrowitz et al. in 1992 (45) reported a clinical trial of automatic closed loop control of the IABP. They used a rule-based algorithm to adjust the time of inflation and deflation. Its safety was verified in anaesthetized dogs and it was then tested on 10 human patients. Their aims were for inflation to occur at dichrotic notch and for deflation to overlap the first half of ventricular ejection. They were successful in 99 and 100% of recordings respectively. Eight of the patients survived. Neither of the two deaths was attributed to the controller.

Sakamoto et al. in 1995 (46) developed a new algorithm to cope with atrial fibrillation, the most unpredictable cardiac arrhythmia sometimes described as irregularly irregular. The aim was for inflation to occur at the dichrotic notch. They were able to predict the time of arrival of the dichrotic notch from mathematical analysis of the R–R interval from the previous 60 heartbeats. Deflation occurred at the R wave. It was tested on ECG recordings from real patients and performed better than conventional timing.

FUTURE DEVELOPMENTS

The use of IABP therapy for preoperative support and as an adjunct to coronary angioplasty or bypass and off-bypass cardiac surgery is likely to increase although larger studies are required to identify which patients will benefit most. The role of the IABP in left ventricular failure is likely to decrease with the increasing use of left ventricular assist devices.

Manufacturers recent research and development has focused on: Improved automatic control algorithms that better cope with alterations in heart rate and rhythm and adjust inflation and deflation times to optimize cardiac assistance. Better catheter designs that cause fewer vascular complications and permit more rapid movement of the shuttle gas.

BIBLIOGRAPHY

1. Harken DE. Presentation at the International College of Cardiology, Brussels; 1958.
2. Clauss RH, et al. Assisted circulation. I. The arterial counterpulsator. J Thorac Cardiovasc Surg 1961;41:447.
3. Kantrowitz A, McKinnen WMP. Experimental use of diaphragm as experimental myocardium. Surg Forum 1958;9:266.
4. Moulopoulos SD, Topaz S, Kolff WJ. Diastolic balloon pumping (with carbon dioxide) in the aorta. A mechanical assistance to the failing circulation. Am Heart J 1962;63:669.
5. Clauss RH, Missier P, Reed GE, Tice D. Assisted circulation by counterpulsation with intra-aortic balloon: Methods and effects. Proceeding of the 4th ACEMB; 1962.
6. Kantrowitz A, et al. Initial clinical experience with intraaortic balloon pumping in cardiogenic shock. J Am Med Ass 1968;203:113.
7. Dennis C, Moreno JR, Hall DP. Studies on external counterpulsation as a potential measure for acute left heart failure. Trans Am Soc Artif Intern Organs 1963;9:186.
8. Kantrowitz A, et al. Initial clinical experience with a new permanent mechanical auxiliary ventricle: The dynamic aortic patch. Trans Am Soc Artif Intern Organs 1972;18(0):159–167, 179.
9. Kantrowitz A, et al. Jr. Initial clinical experience with intraaortic balloon pumping in cardiogenic shock. JAMA 1968;203(2):113–118.
10. Kantrowitz A. Origins of intraaortic balloon pumping. Ann Thoracic Surg 1990;50(4):672–674.
11. Freed PS, Wasfie T, Zado B, Kantrowitz A. Intraaortic balloon pumping for prolonged circulatory support. Am J Cardiol 1988;61(8):554–557.
12. Lund O, et al. Intraaortic balloon pumping in the treatment of low cardiac output following open heart surgery–immediate results and long-term prognosis. Thoracic Cardiovascular Surg 1988;36(6):332–337.
13. Meharwal ZS, Trehan N.Vascular complications of intra-aortic balloon insertion in patients undergoing coronary reavscularization: Analysis of 911 cases.[See comment]. Eur J Cardio-Thoracic Surg 2002;21(4):741–747.
14. Ferguson JJ. 3rd, et al. The current practice of intra-aortic balloon counterpulsation: Results from the benchmark registry. J Am College Cardiol 2001;38(5):1456–1462.
15. Dietl CA, et al. Efficacy and cost-effectiveness of preoperative iabp in patients with ejection fraction of 0.25 Or less. Ann Thorac Sur 1996;62(2):401–408.
16. Christenson JT, Simonet F, Badel P, Schmuziger M. Evaluation of preoperative intra-aortic balloon pump support in high risk coronary patients. Eur J Cardio-Thoracic Sur 1997;11(6): 1097–1103.
17. Suzuki T, et al. Usefulness of preoperative intraaortic balloon pump therapy during off-pump coronary artery bypass grafting in high-risk patients. Ann Thoracic Surg 2004;77 (6):2056–2059.
18. Brodie BR, Stuckey TD, Hansen C, Muncy D. Intra-aortic balloon counterpulsation before primary percutaneous transluminal coronary angioplasty reduces catheterization laboratory events in high-risk patients with acute myocardial infarction. Am J Cardiol 1999;84(1):18–23.
19. Ishihara M, Sato H, Tateishi H, Uchida T, Dote K. Intraaortic balloon pumping as the postangioplasty strategy in acute myocardial infarction. Am Heart J 1991;122(2):385–389.
20. Ohman EM, et al. Use of aortic counterpulsation to improve sustained coronary artery patency during acute myocardial infarction. Results of a randomized trial. The randomized iabp study group Circulation 1994;90(2):792–799.
21. Stone GW, et al. A prospective, randomized evaluation of prophylactic intraaortic balloon counterpulsation in high risk patients with acute myocardial infarction treated with primary angioplasty. Second primary angioplasty in myocardial infarction (pami-ii) trial investigators. J Am College Cardiol 1997;29(7):1459–1467.
22. van't Hof AW, et al. A randomized comparison of intra-aortic balloon pumping after primary coronary angioplasty in high risk patients with acute myocardial infarction. [See comment]. Eur Heart J 1999;20(9):659–665.
23. Sanborn TA, et al. Impact of thrombolysis, intra-aortic balloon pump counterpulsation, and their combination in cardiogenic shock complicating acute myocardial infarction: A report from the shock trial registry. Should we emergently revascularize occluded coronaries for cardiogenic shock? J Am College Cardiol 2000;36(3 Suppl. A):1123–1129.
24. Minich LL, et al. Intra-aortic balloon pumping in children with dilated cardiomyopathy as a bridge to transplantation. J Heart Lung Transplant 2001;20(7):750–754.
25. Pantalos GM, et al. Estimation of timing errors for the intraaortic balloon pump use in pediatric patients. ASAIO Journal 1999;45(3):166–671.
26. Akomea-Agyin C, et al. Intraaortic balloon pumping in children. Ann Thoracic Surg 1999;67(5):1415–1420.
27. Pinkney KA, et al. Current results with intraaortic balloon pumping in infants and children. Ann Thoracic Surg 2002;73(3):887–891.
28. Kantrowitz A, et al. Intraaortic balloon pumping 1967 through 1982: analysis of complications in 733 patients. Am J Cardiol 1986;57(11):976–983.
29. Miller JS, Dodson TF, Salam AA, Smith RB3rd. Vascular complications following intra-aortic balloon pump insertion. Am Surg 1992;58(4):232–238.
30. Macoviak J, et al. The intraaortic balloon pump: an analysis of five years' experience. Ann Thoracic Surg 1980;29(5):451–458.

31. Vonderheide RH, Thadhani R, Kuter DJ. Association of thrombocytopenia with the use of intra-aortic balloon pumps. Am J Med 1998;105(1):27–32.
32. Jaron D, Ohley W, Kuklinski W. Efficacy of counterpulsation: model and experiment. Trans Am Soc Artificial Inter Organs 1979;25:372–377.
33. Jaron D, Moore TW, He P. Theoretical considerations regarding the optimization of cardiac assistance by intraaortic balloon pumping. IEEE Trans Biome Eng 1983;30(3):177–185.
34. Jaron D, Moore TW, He P. Control of intraaortic balloon pumping: theory and guidelines for clinical applications. Ann Biomed Eng 1985;13(2):155–175.
35. Niederer P, Schilt W. Experimental and theoretical modelling of intra-aortic balloon pump operation. Med Biol Eng Comput 1988;26(2):167–174.
36. Barnea O, Smith B, Moore TW, Jaron D. Simulation and optimization of intra-aortic balloon pumping. Proceedings. Computers in Cardiology (Cat. No.89CH2932-2). Los Alamitos (CA): IEEE Computer Society Press; 1990. p 237–240.
37. Sakamoto T, et al. Effects of timing on ventriculoarterial coupling and mechanical efficiency during intraaortic balloon pumping. ASAIO Journal 1995;41(3):M580–MM583.
38. Weller PR, Morrow DR, LeFèvre JE. Evolution of a fuzzy controller for the intra-aortic balloon pump. Proceedings of 2nd European Medical and Biological Engineering Conference, EMBEC'02, Vienna, Austria, Hutten H, Krosl P. editors. ISBN 3-901351-62-0;4–8. December 2002; p 1588–1589.
39. Kane GR, Clark JW, Bourland HM, Hartley CJ. Automatic control of intra-aortic balloon pumping. Trans Am Soc Artif Int Organs 1971;17:148.
40. Martin PJ, Jaron D. New controller for in-series cardiac-assist devices. Med Biol Eng Computing 1978;16(3):243–249.
41. Barnea O, Smith B, Moore TW, Jaron D. An optimal control algorithm for intra-aortic balloon pumping. Images of the Twenty-First Century. Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (Cat. No.89CH2770-6). New York: Vol. 5. IEEE; 1989. p 1419–1420.
42. Barnea O, et al. Optimal controller for intraaortic balloon pumping. IEEE Trans. Biomed Eng 1992;39(6):629–634.
43. Smith B, Barnea O, Moore TW, Jaron D. An algorithm for optimal control of the intra-aortic balloon pump. Proceedings of the Fifteenth Annual Northeast Bioengineering Conference (Cat. No.89-CH2689-8). New York; IEEE. 1989; p 75–76.
44. Zelano JA, Li JK, Welkowitz W. A closed-loop control scheme for intraaortic balloon pumping. IEEE Trans Biomed Eng 1990;37(2):182–192.
45. Kantrowitz A, et al. Initial clinical trial of a closed loop, fully automatic intra-aortic balloon pump. ASAIO Journal 1992; 38(3):M617–M621.
46. Sakamoto T, Arai H, Maruyama T, Suzuki A. New algorithm of intra aortic balloon pumping in patients with atrial fibrillation. ASAIO Journal 1995;41(1):79–83.

Further Reading

Goldberger M, Tabak SW, Shah PK. Clinical experience with intra-aortic balloon counterpulsation in 112 consecutive patients. *Am Heart J* 1986;111(3):497–502.

McEnany MT, et al. Clinical experience with intraaortic balloon pump support in 728 patients. *Circulation* 1978;58(3 Pt 2):I124–I132.

See also ARTERIES, ELASTIC PROPERTIES OF; HEMODYNAMICS; VASCULAR GRAFT PROSTHESIS.

INTRACRANIAL PRESSURE MONITORING. See MONITORING, INTRACRANIAL PRESSURE.

INTRAOCULAR LENSES. See LENSES, INTRAOCULAR.

INTRAOPERATIVE RADIOTHERAPY. See RADIOTHERAPY, INTERAOPERATIVE.

INTRAUTERINE DEVICES (IUDS). See CONTRACEPTIVE DEVICES.

INTRAUTERINE SURGICAL TECHNIQUES

JOSEPH P. BRUNER
Department of Obstetrics and Gynecology
Nashville, Tennessee

R. DOUGLAS WILSON
N. SCOTT ADZICK
University of Pennsylvania
Philadelphia, Pennsylvania

LOUISE WILKINS-HAUG
AUDREY C. MARSHALL
Women's and Children's Hospital
Boston, Massachusetts

RUBEN A. QUINTERO
Florida Institute for Fetal Diagnosis and Therapy
Tampa, Florida

M. YAMAMOTO
Y. VILLE
University of Paris
Paris, France

ANTHONY JOHNSON
University of North Carolina
Chapel Hill, North Carolina

JULIE S. MOLDENHAUER
Wayne State University
Detroit, Michigan

INTRODUCTION

Intrauterine surgery of the fetus or fetal adnexae is spreading rapidly throughout the world. In a broad sense, intrauterine surgery includes any procedure in which a medical device is purposely placed within the uterine cavity. For most physicians, however, the concept of intrauterine surgery excludes such commonly performed procedures as amniocentesis and chorionic villous sampling. Rather, the term usually refers to techniques requiring specialized knowledge, experience, and most especially, instrumentation. Procedures most commonly mentioned in discourses on intrauterine surgery include those specifically developed for the treatment of such serious anatomic defects as congenital cystic adenomatoid malformation (CCAM), sacrococcygeal teratoma (SCT), lower urinary tract obstruction (LUTO), myelomeningocele, aortic or pulmonic stenosis, gastroschisis, iatrogenic amniorhexis, twin-to-twin transfusion syndrome (TTTS), and twins discordant for severe anomalies. As no one person in the world is an expert in every one of these procedures, the remainder of this chapter is divided into individual sections, each authored by someone widely recognized as a leader in the treatment of that particular anomaly.

CONGENITAL CYSTIC ADENOMATOID MALFORMATION AND SACROCOCCYGEAL TERATOMA

Most prenatally diagnosed malformations are managed by appropriate medical and surgical evaluation and treatment following planned delivery near term, with some cases requiring transfer of the mother to a tertiary referral center with obstetrics, maternal fetal medicine, medical genetics, neonatology, and pediatric surgery subspecialties. Certain anatomic abnormalities can have significant fetal developmental consequences, with emergency *in utero* therapy being required due to gestational age and mortality risks for the fetus. Open maternal fetal surgery poses additional risks to the mother. Maternal fetal surgery (1–4) should not be attempted until (1) the natural history of the fetal disease is established by following up on untreated cases, (2) selection criteria for cases requiring intervention are developed, (3) pathophysiology of the fetal disorder and its correction are defined in fetal animal models, and (4) hysterotomy and fetal surgery can be performed without undue risk to the mother and her reproductive potential.

Congenital cystic adenomatoid malformation of the lung (CCAM) (5) is a rare lesion characterized by a multicystic mass of pulmonary tissue with proliferation of bronchial structures. CCAM is slightly more common in males and is unilobar in 80–95% of cases. CCAMs derive their arterial blood supply from the normal pulmonary circulation. CCAMs are divided clinically into cystic and solid lesions, but have been divided traditionally into three types based on their pathological characteristics. Type 1 CCAM lesions account for 50% of postnatal CCAM cases and consist of single or multiple cysts lined by ciliated pseudostratified epithelium (5). These cysts are usually 3–10 cm in size and 1–4 in number (5). Type II CCAM lesions account for 40% of postnatal cases of CCAM and consist of more numerous cysts of smaller diameter, usually less than 1 cm. They are lined by ciliated, cuboidal, or columnar epithelium (5). Type III CCAM lesions account for only 10% of CCAM cases and are usually large, homogenous, microcystic masses that cause mediastinal shift. These lesions have bronchiolar-like structures lined by ciliated cuboidal epithelium separated by masses of alveolar-sized structures lined by nonciliated cuboidal epithelium (5). Prognosis in Type III CCAM is related to size.

Sacrococcygeal teratoma (SCT) (6) is defined as a neoplasm composed of tissues from either all three germ layers or multiple foreign tissues lacking an organ specificity. SCT is thought to develop from a totipotent somatic cell originating in Hensen's node. SCT has been classified by the relative amounts of presacral and external tumor present, with Type I completely external with no presacral component, Type II external component with internal pelvic component, Type III external component with internal component extending into the abdomen, and Type IV completely internal with no external component (7). A Type I SCT is evident at birth and is usually easily resected and has a low malignant potential (6). Type II and III SCTs are recognized at birth, but resection may be difficult requiring an anterior and posterior approach (6). A Type IV SCT can have a delayed diagnosis until symptomatic at a later age (6). SCT is one of the most common tumors in newborns and has an incidence of 1 per 35,000 to 40,000 live births (6).

Evaluation of the fetal status for CCAM and SCT requires multiple imaging and functional techniques (8–10), including fetal ultrasound, fetal MRI, and fetal echocardiogram with arterial and venous Doppler assessments (umbilical artery, umbilical vein, ductus venosus). Measurements include combined cardiac output, cardiothoracic ratio, descending aortic blood flow, inferior vena cava diameter, placental thickness, umbilical artery systolic to diastolic Doppler ratio, and amniotic fluid index. Presence of ascites, pleural or pericardial effusion, and skin or scalp edema are important markers for the extent of fetal hydrops and its overall effect on fetal stability. Specific ultrasound imaging of the CCAM and SCT looks for the percentage of cystic and solid components in the tumors as well as an overall mass volume (cc) estimate (AP cm × transverse cm × height cm × 0.52). The SCT consistency and size can be reflected directly in the combined cardiac output and amount of vascular shunting. The CCAM overall size can cause mediastinal shift with cardiac dysfunction and pulmonary deformation. Validated ratio of CCAM/head circumference (CVR) can be used for prognosis and follow-up planning (10). The specific lobar location for the CCAM may have a differential impact on cardiac function. The development of fetal hydrops is due mainly to cardiac dysfunction secondary to compression.

The physiologic changes required in the fetal status to move from expectant management to open maternal fetal surgery is generally dictated by fetal (gestational age and extent of fetal hydrops) and maternal factors (8–10). Criteria for consideration of maternal fetal surgery for CCAM resection (fetal lobectomy) require the absence of maternal risk factors for anesthesia and surgery, a singleton pregnancy with a normal karyotype (amniocentesis, chorionic villus sampling, or percutaneous umbilical blood sampling), no other anatomical abnormalities beyond the associated hydrops, gestational age of 21–31 weeks, and massive multicystic or predominantly solid CCAM (CVR > 1.6) (8–10). In selected cases, the failure of *in utero* therapy techniques, such as thoracoamniotic shunting or cyst aspiration for the large Type I lesions, to reverse the fetal hydrops would be required. Criteria for consideration of maternal fetal surgery for debulking of a SCT require the absence of maternal risk factors for anesthesia and surgery, a singleton pregnancy with a normal karyotype, the absence of significant associated anomalies, evidence of impending high output cardiac failure, gestational age of 21–30 weeks, and favorable SCT anatomy classification (Type I or II) (9).

The technique for maternal hysterotomy to allow access to the fetus has been well described and has evolved over 25 years of experimental and clinical work (1,8,9). The uterus is exposed through a maternal low transverse abdominal incision. If a posterior placenta is present, superior and inferior subcutaneous flaps are raised and a vertical midline fascial incision is made to expose the uterus for a convenient anterior hysterotomy with the uterus remaining in the abdomen. Conversely, the presence of an anterior placenta necessitates the division of the rectus muscles so the uterus can be tilted out of the abdomen for a posterior hysterotomy. A large abdominal ring retractor (Turner–Warwick) is used to maintain exposure and prevent lateral compression of the uterine vessels. Sterile interoperative ultrasound is used to delineate the fetal position and placental location. The edge of the placenta is marked under sonographic guidance using electrocautery or a marking pen. The position and orientation of the hysterotomy is planned to stay parallel to and at least 6 cm from the placental edge but still allow exposure of the appropriate fetal anatomy. The hysterotomy is facilitated by the placement of two large monofilament sutures (PDS II 1 Ethicon; Somerville, NJ) parallel to the intended incision site and through the full thickness of the uterine wall and membranes under sonographic guidance. The electrocautery is used to incise the myometrium between the two stay sutures down to the level of the amniotic membranes. A uterine stapler device (US Surgical Corporation; Norwalk, CT) with absorbable Lactomer staples is then directly introduced through the point of fixation and into the amniotic cavity by using a piercing attachment on the lower limb of the stapler. The stapler is fired, thereby anchoring the amniotic membranes (chorion, amnion) to the uterine wall creating a hemostatic hysterotomy. Careful evaluation for the membrane adhesion status and for any myometrial bleeding sites is undertaken. If required, interrupted PDS sutures are used to control bleeding and membrane separation. The fetus and the internal uterine cavity are continually bathed in warmed lactated Ringers at 38–40°C using a level I warming pump connected to a red rubber catheter that is placed in the uterine cavity through the hysterotomy.

For CCAM resection (1,8,11), once the appropriate fetal area is visualized in the hysterotomy site, the fetal arm is brought out for pulse oximeter monitoring, IV access, and fetal position control. Intraoperative fetal echocardiography is used throughout to monitor cardiac function. The fetal chest is entered by a fifth intercostal space thoracotomy. The lesion usually decompresses out through the thoracotomy wound consistent with the increase in the thoracic pressure from the mass (8). Using techniques initially developed on experimental animals, the appropriate pulmonary lobes containing the lesion are resected (1,11). Fetal resuscitation is performed if needed through intravenous administration of crystalloid, blood, and code-blue medications with fetal echocardiography providing functional information. The fetal thoracotomy is closed and the fetal arm is returned to the uterus.

The technique for debulking of an external fetal SCT has been described in detail previously (1,9,12,13). The fetal foot is used for pulse oximeter monitoring and IV access with intraoperative echocardiography. The fetal SCT is exposed and a Hagar dilator is placed in the rectum. Fetal skin is incised circumferentially around the base of the tumor and a tourniquet is applied to constrict blood flow. The tumor is debulked externally, usually with a 90 mm thick tissue stapler (US Surgical Corporation; Norwalk, CT). The objective of the fetal SCT resection is to occlude the tumor vascular supply and remove the low resistance tumor vascular bed from the fetal circulation. No attempt is made to dissect the intrapelvic component of the tumor or to remove the coccyx (done with a second procedure after birth). Fetal resuscitation is performed if needed through intravenous administration of crystalloid, blood, and code-blue medications with fetal echocardiography providing functional information. The fetal sacral wound is closed.

Repair of the hysterotomy after fetal surgery (1–4) uses a water-tight two-layered uterine closure, with interrupted full thickness stay sutures placed first and untied using PDS II 1 (Ethicon; Somerville, NJ), and the uterus is then closed with a running continuous stitch PDSII 0 (Ethicon; Somerville, NJ) including the chorion-amnion membrane layer. The interrupted stay sutures are then tied after the amniotic fluid volume has been corrected with warm lactated Ringers through a red rubber catheter and volume confirmed by ultrasound visualization. The omentum is sutured in place over the hysterotomy closure to help seal the hysterotomy site with vascularized tissue and to prevent bowel adherence to the site, especially when a posterior hysterotomy is performed. The maternal laparotomy incision is closed in layers. It is important to use a subcuticular skin closure covered with a transparent dressing so that monitoring devices can be placed on the maternal abdomen postoperatively.

In some specific cases, when the CCAM lesion is not resected *in utero*, it continues to be a large space-occupying lesion with mediastinal shift. Thus, it might be anticipated that respiratory compromise will be present at birth, the delivery may be facilitated with an EXIT procedure (*ex utero* intrapartum therapy) (14). Uterine relaxation is maintained by high concentration inhalational anesthetics, with additional tocolysis if necessary. The EXIT requires only the head and chest to be initially delivered through, preferably, a low transverse hysterotomy wound thereby preserving uterine volume with the lower fetal body and continuous warmed lactated Ringers infusion to prevent cord compression. These maneuvers preserve the uterine-placental circulation and continue placental gas exchange. The EXIT procedure can be done through an anterior or posterior hysterotomy, but its location in the uterus may require that all future pregnancies be delivered by cesarean section with no trial of labor if a low anterior transverse location is not available.

All future pregnancies following maternal hysterotomy for maternal-fetal surgery require cesarean section at term with no trial of labor. Maternal obstetrical risks in a subsequent pregnancy are similar to risks following for a classic cesarean section (15).

LOWER URINARY TRACT OBSTRUCTION

The diagnosis and treatment of fetal lower urinary tract obstruction (LUTO) requires knowledge of the differential diagnosis and the natural history of the condition, a thorough understanding of the criteria for therapy, and management expertise. Fetal LUTO is one of the most commonly diagnosed birth defects. Untreated, and depending on the level of the obstruction, it may lead to hydronephrosis, renal dysplasia, pulmonary hypoplasia, and perinatal death (16,17). The prognosis depends on the extent of preexisting renal damage and the effectiveness of therapy. Treatment with fetal urinary diversion procedures is aimed at preventing renal damage and pulmonary hypoplasia (18–20).

Obstruction to urine flow has been shown in animal models to result in hydronephrosis and renal dysplasia (21). Release of the obstruction is associated with no or variable renal damage depending on the timing of the release or the creation of the defect (21,22). Pulmonary hypoplasia is another major potential complication of fetuses with obstructive uropathy (23). The association probably results from the attendant oligohydramnios.

Urethral obstruction may result from posterior urethral valves (PUV), anterior urethral valves, megalourethra, urethral duplications, urethral atresia, obstructive ureterocele, or cloacal dysgenesis. Posterior urethral valves (PUV), first described by Young et al. (24), constitute the most common cause of lower urinary tract obstruction in male neonates, with an incidence of 1:8000 to 1:25,000 livebirths (25). The lesions occur only in males because the female counterpart of the verumontanum, from which the valves originate, is the hymen.

In utero therapy is usually limited to fetuses with bladder outlet obstruction. Fetuses with unilateral obstruction are not typically considered candidates for *in utero* therapy, regardless of the magnitude of the obstruction or renal findings. In these patients, the risk/benefit ratio of *in utero* intervention favors expectant management, even if it means loss of the affected renal unit.

Fetal renal function may be assessed by analysis of fetal urinary parameters via vesicocentesis. Patients are considered candidates for *in utero* therapy if fetal urinary parameters are below the threshold for renal cystic dysplasia. If the values are above the threshold, therapy should not be offered.

The application of selection criteria in patients with fetal LUTO for possible *in utero* therapy results in a significant attrition rate. Disqualification from therapy may result both from “too healthy” or “too sick” conditions. Examples of too healthy conditions include normal amniotic fluid volume or suggestion of nonobstructive dilatation of the urinary tract. Examples of too sick conditions include sonographic evidence of renal cystic dysplasia, abnormal fetal urinary parameters, abnormal karyotype, or the presence of associated major congenital anomalies. Of 90 patients referred to the Florida Institute for Fetal Diagnosis and Therapy from October 1996 to October 2003, more than one-half were disqualified from therapy from single or overlapping conditions.

Percutaneous ultrasound-guided vesicoamniotic shunting of fetuses with LUTO began in the early 1980s (16,19,23). The goal of therapy is to avoid development of pulmonary hypoplasia from the attendant oligohydramnios as well as to preserve renal function. Fetal bladder shunting should be offered only to patients without sonographic or biochemical evidence consistent with renal cystic dysplasia, normal karyotype, and lack of associated major congenital anomalies.

The procedure can be performed under local, regional, or general anesthesia. A minimal skin incision is made. Ultrasound is used to identify the ideal site of entry into the fetal bladder, below the level of the umbilicus. Color Doppler ultrasonography is used to identify the umbilical vessels around the distended bladder and avoid them. Under ultrasound guidance, the trocar is directed through the maternal tissues and up to the fetal skin. Fetal analgesia is achieved with pancuronium 0.2 mg/kg and fentanyl 10 mcg/kg. The trocar stylet is used to enter the fetal bladder with a sharp, swift, and controlled maneuver. If a prior vesicocentesis had been performed, it is advisable to obtain a sample of fetal urine for microbiological purposes to rule out preexisting infection. A sample of fetal urine is sent for further biochemical testing. Placement of the double-pigtail catheter is monitored with ultrasound. After the distal loop is deployed in the bladder, the trocar is retrieved to the level of the bladder wall. A small amount of the straight portion of the catheter may be advanced into the bladder to avoid retracting the distal loop into the bladder wall. The trocar shaft is retrieved slowly while simultaneously maintaining pressure on the catheter to deploy the straight portion within the bladder wall and fetal skin. Once the shaft of the trocar reaches the fetal skin, entrance of the catheter, including the proximal loop, can be safely deployed. If complete anhydramnios is present prior to insertion of the catheter, it is advantageous to attempt an amnioinfusion with an 18 gauge needle prior to shunting to create the space for deployment of the proximal loop. Amnioinfusion is aimed at preventing misplacement of the proximal loop within the myometrium and fetal membranes.

Despite adequate placement, malfunction of vesicoamniotic shunting may occur up to 60% of the time (26). The shunt may pull from the skin into the fetal abdomen, resulting in iatrogenic ascites, or out of the fetal bladder, with no further drainage of urine. The shunt may pull out of the fetus altogether as well. Replacement of the shunt is associated with an additive risk of fetal demise, chorioamnionitis, premature rupture of membranes, and miscarriage or preterm delivery, for a total perinatal loss rate of approximately 5% per instance.

In 1995, we proposed the use of endoscopy to assess the fetal bladder for diagnostic and surgical purposes (27,28). Endoscopic visualization of the fetal bladder with a larger endoscope can be justified during vesicoamniotic shunting. Currently, we use a 3 mm or a 3.9 mm trocar with a 2.7 mm or 3.3 mm diagnostic or operating endoscope. This diameter is slightly larger than the 14 gauge (approximately 2.1 mm) needle used for the insertion of the double-pigtail catheter. Access to the fetal bladder allows remarkable evaluation of the bladder, ureteral orifices, and urethra as well as the opportunity to perform surgical procedures.

In normal fetuses, the urethra is not dilated, appearing as a small hole within the bladder. In patients with a true urethral obstruction, endoscopy will show a variable dilatation of the urethra at the level of the bladder neck. The urethra is located using a 25° or a 70° diagnostic rigid endoscope. Alternatively, a flexible/steerable endoscope may be used. The anatomical landmarks to identify at this level include the verumontanum and the urethral valves. The diagnostic endoscope is then exchanged for a rigid operating endoscope. A 600 μm YAG-laser fiber is passed through the operating channel of the endoscope, and then ablated using 5–10 w and 0.2 s pulses in successive steps. The fiber is placed as anterior and medial as possible. It is not necessary to evaporate the entire valvular tissue. Instead, only a few defects to either side of the midline are necessary to establish urethral patency (27,29). The dilated urethra may collapse intraoperatively once patency is re-established, which may obscure the field of view and require frequent instillation of saline to the side port of the trocar to distend it. Color Doppler may also be used to document fetal urination through the penis.

A urethrorectal fistula may occur from thermal damage beyond the posterior wall of the urethra into the perirectal space. To avoid this complication, only 5–10 w of energy in short bursts should be used while ablating the valves.

The management of fetuses with lower obstructive uropathy continues to be one of the most challenging subjects in fetal therapy. The difficulties include establishing the correct differential diagnosis, accurately predicting subsequent renal function, and providing the best treatment.

MYELOMENINGOCELE

Myelomeningocele results from the failure of caudal neural tube closure during the fourth week of gestation. The lesion is characterized by protrusion of the meninges through a midline bony defect of the spine, forming a sac containing cerebrospinal fluid and dysplastic neural tissue. Affected infants exhibit varying degrees of somatosensory loss, neurogenic sphincter dysfunction, paresis, and skeletal deformities (30). Virtually all such infants also have the Chiari II malformation, and up to 95% develop hydrocephalus (31). Although myelomeningocele is not a lethal disorder, the neurologic sequelae are progressive, and worsen until the lesion is closed. Observational and cohort studies have demonstrated improvement of the Chiari II malformation (32,33), decreased hydrocephalus (34), and improved lower extremity function after intrauterine repair of myelomeningocele (35).

On the day of surgery, the pregnant patient is taken to a standard obstetrical operating room. An epidural catheter is placed and, after induction of general endotracheal anesthesia, she is prepared as if for a cesarean section. Many of the general anesthetic agents cross the placenta and provide analgesia for the fetus, and the epidural catheter enables the administration of continuous postoperative analgesics if needed. The gravid uterus is exposed with a Pfannenstiel incision and exteriorized. The uterine contents are then mapped with a sterile ultrasound transducer, and the location of the fetus and the placenta are determined. Initial uterine entry is obtained with a specialized trocar developed at Vanderbilt University Medical Center (Cook Incorporated; Bloomington, IN). The Tulipan–Bruner trocar consists of a tapered central introducer covered by a peel-away Teflon sheath. Use of this trocar has demonstrated to reduce operative time and blood loss while providing atraumatic entry into the uterine cavity (36). Two through-and-through chromic sutures are passed through the uterine wall and membranes on either side of the selected entry point. The introducer is then passed into the uterine cavity under direct ultrasonographic guidance using a modified Seldinger technique. The central introducer is then removed, leaving only the trocar sheath. Excess amniotic fluid may be aspirated and stored in sterile, warm syringes. The footplate of a U.S. surgical CS-57 autostapling device (United States Surgical Corporation; Norwalk, CT) is then inserted through the peel-away sheath, and the sheath is removed, leaving the stapler in proper position. When activated, the stapler creates a 6–8 cm uterine incision. At the same time, all the layers of the uterine wall are held together, much like the binding of a book.

The fetus is directly visualized and manually positioned within the uterus so that the myelomeningocele sac is located in the center of the hysterotomy. Proper position is maintained by grasping the fetal head and trunk through the flaccid uterine wall. During the procedure, the fetal heart rate is monitored by continuous ultrasonographic visualization.

The myelomeningocele is closed in routine neurosurgical fashion. Approximately 20% of patients will not have a well-formed myelomeningocele sac, but a crater-like lesion termed myeloschisis. As fetuses with myeloschisis have less viable skin for closure, it may be necessary to use bilateral vertical relaxing incisions in the flanks to create bipedicular flaps that can be advanced and closed over the dural sac. The resulting full-thickness cutaneous defects are covered with cadaveric skin (37).

After repair of the spina bifida lesion, the uterus is closed in layers using #1 PDS sutures. The first layer incorporates the absorbable polyglycolic acid staples left by the autostapling device. As the last stitches of this layer are placed, the reserved amniotic fluid or physiologic crystalloid solution, mixed with 500 mg of nafcillin or an equivalent dosage of an antibiotic effective against Staphylococcus species, is replaced in the uterus. The sterile, warm fluid is added until the uterine turgor, as determined by manual palpation, is restored to the preoperative level, which is followed by an imbricating layer. A sheet of Interceed absorbable adhesion barrier (Johnson & Johnson Medical, Inc.; Arlington, TX) or omentum is attached over the incision to prevent adhesion formation. The uterus is returned to the abdomen. The fascial layer is closed in routine fashion, and the dermis closed with a running subcuticular suture or staples. The fetus is monitored postoperatively using continuous electronic fetal monitoring (EFM) and intermittent transabdominal ultrasonography.

Postoperative uterine contractions are monitored using continuous EFM. Uterine contractions are initially controlled with intravenous magnesium sulfate and oral or rectal indomethacin, and subsequently with subcutaneous terbutaline or oral nifedipine, supplemented by indomethacin as needed. Patients are monitored with weekly transabdominal ultrasonographic examinations. Delivery of each child is accomplished via standard cesarean section. Although the same abdominal incision is used for the cesarean section as for the fetal surgery, the fetus is preferably delivered via a lower uterine segment incision.

The uterus and abdominal incisions are closed in routine fashion.

VALVULOPLASTY

Severe aortic stenosis in midgestation may lead to left ventricular myocardial damage and can ultimately result in hypoplastic left heart syndrome. Paradoxically, in these fetuses, which are likely to progress to HLHS, the left ventricle initially appears normal in size, or even enlarged, in the setting of left ventricular systolic dysfunction. As gestation progresses, diminished flow through the diseased left ventricle leads to decreased flow, and the ventricle experiences growth arrest, resulting in left heart hypoplasia at birth. Early relief of fetal aortic stenosis may preserve left heart function and growth potential by maintaining flow through the developing chamber. To this end, a number of operators have developed techniques to perform fetal aortic valvuloplasty in second trimester fetuses.

The mother is placed under general anesthesia in a supine position with left lateral uterine displacement. Transabdominal ultrasound imaging and external manipulation are employed to achieve ideal fetal position. In this position, a line of approach from the anterior abdominal surface traverses the apex of the fetal left ventricle (LV), paralleling the LV outflow tract, and crossing the valve into the ascending aorta. The fetus is given intramuscular anesthetic and muscle relaxant prior to catheterization. If unable to position the fetus using external maneuvers, the operators perform a limited laparotomy to enable direct uterine manipulation and transuterine imaging.

A low profile, over-the-wire coronary angioplasty catheter is chosen with a balloon diameter based on the measurement of the aortic annulus, using a balloon:annulus ratio of 1:2. The balloon catheter is mounted on a floppy-tipped guidewire, with 3 cm of distal wire exposed. The wire/catheter assembly is then advanced through the 19G 12 cm stainless-steel introducer cannula until the balloon emerges. Affixing a visible and palpable marker on the proximal catheter shaft allows the operator to reproduce this balloon/cannula relationship during the procedure without relying wholly on the ultrasound imaging.

The introducer is advanced through the fetal chest wall and to the LV epicardium under ultrasound guidance. The LV is entered with the introducer, and the obturator removed with the tip of the cannula just below the aortic valve (Fig. 1). Blood return through the cannula confirms an intracavitary position.

The wire/catheter assembly is passed through the cannula, and the tip of the wire is identified as it emerges. While maintaining imaging of the aortic valve and ascending aorta, the precurved wire tip is manipulated to probe for the valve. Valve passage, confirmed echocardiographically by imaging the wire in the ascending aorta, is followed by catheter insertion to the premarked depth. The balloon is then inflated, by hand or by pressure gauge, to a pressure at which it achieves the intended balloon:annulus ratio. Upon completion of the dilation, the entire apparatus is removed from the fetus.

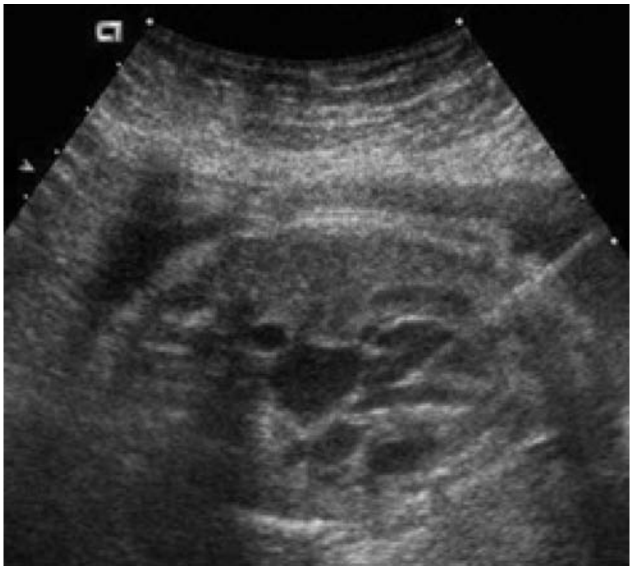


Figure 1. Transabdominal ultrasound imaging during introducer cannula insertion into dilated fetal left ventricle. The tip of the introducer is positioned in the left ventricular outflow tract directed at the stenotic aortic valve.

When first reported in the 1990s, fetal aortic valve dilation was performed with minimal technical success (38). Using the technique described above, technical success rates are now over 80% in fetuses between 21 and 26 weeks (39). As the technical aspects of the procedure continue to be refined, and safety is established, issues of patient selection will become the major focus of ongoing research. Anatomic and physiologic variables predicting left ventricular normalization following successful fetal aortic valvuloplasty remain poorly understood.

AMNIOEXCHANGE

Gastroschisis is a paraumbilical defect of the anterior abdominal wall associated with intrauterine evisceration of the fetal abdominal organs. The incidence of gastroschisis is approximately 1:4000 births, with a 1:1 male:female ratio. Most cases are sporadic and aneuploidy is uncommon.

Gastroschisis is characterized by a full-thickness defect of the abdominal wall, usually located to the right of the umbilical cord, which has a normal insertion. The defect in the abdominal wall is generally quite small (3–5 cm). The herniated organs include mainly bowel loops, although, in rare cases, the spleen and liver may be involved. Intestinal atresias and other gastrointestinal disruptions are found in as many as 15% of cases, and malrotation is also universal.

Although the prognosis is excellent with an ultimate survival of greater than 90%, many factors may jeopardize the outcome of these infants. A chronic aseptic amniotic fluid peritonitis (perivisceritis) often occurs. The herniated organs become covered by an inflammatory peel in the third trimester, resulting from chemical irritation by exposure to digestive enzymes in the amniotic fluid. Thickening, edema, and matting together of the intestines occurs in these cases, and may result in a secondary ischemic injury to the bowel as the abdominal defect becomes too small. Meconium is frequently found in the amniotic fluid of affected fetuses. Its presence probably reflects intestinal irritation. Intrauterine growth restriction (IUGR) is frequent, occurring in up to 60% of fetuses. Oligohydramnios can occur, and may lead to fetal stress by cord compression. Premature birth is a frequent and still poorly understood complication. At birth, infants have low serum albumin and total protein levels, which probably results from chronic peritonitis.

The obstetrical management of the fetus with gastroschisis is controversial. Some studies have shown no clear benefit of cesarean delivery over vaginal delivery, where as others demonstrate an improved perinatal outcome in infants delivered by elective cesarean section prior to labor. Postoperative infection and delayed total enteral nutrition are the major acute complications of the newborn. Although all neonates with gastroschisis require surgery shortly after birth, repair may be by primary fascial closure, or by delayed fascial closure using temporary coverage with a silastic/Dacron intra-abdominal pouch. The repair as a primary or secondary procedure depends on the degree of chemical peritonitis with matting of the bowel that is present. Delayed intestinal function with poor enteral nutrition is expected in most patients. Central venous access and early total parenteral nutrition are therefore usually required.

In a study by Luton et al. (40), gastroschisis was created at mid gestation in 21 lamb fetuses. Saline was amnioinfused in some fetuses every 10 days until term. Thickness of the bowel muscularis, thickness of the serous fibrosis, and plasma cell infiltration were all significantly improved in amnioexchanged animals when compared with fetal lambs that were not amnioexchanged (40). Histologic analysis of appendices removed from human newborns demonstrated increased fibrosis in those with gastroschisis; after amnioinfusion, the serosa was still edematous, but no inflammation was seen (41). In a pilot study in human pregnancies (42), the same authors investigated the effect of amnioinfusion on the outcome of prenatally diagnosed gastroschisis. Following up on their work showing that an inflammatory response exists in the amniotic fluid of fetuses with gastroschisis, they hypothesized that amniotic fluid exchange would improve the outcomes of prenatally diagnosed cases. The outcome of 10 amnioinfused fetuses with gastroschisis was compared with 10 nonamnioinfused matched controls. Results showed that fetuses undergoing amnioinfusion had a shorter duration of curarization after surgical repair (2.2 ± 1.9 versus 6.8 ± 6.9 days, $+ = 0.019$), a shorter delay before full oral feeding (49.7 ± 21.5 versus 72.3 ± 56.6 days, NS) and a shorter overall length of hospitalization (59.5 ± 19.7 versus 88.5 ± 73.6 days, NS). The authors confirmed their previous data showing that amniotic fluid displays a chronic inflammatory profile, and they speculated that a reduction of the inflammatory response could improve the outcome of human fetuses with gastroschisis (42).

Amnioexchange for treatment of gastroschisis begins after 30 weeks' gestation, and is repeated approximately every two weeks until delivery. A complete obstetrical ultrasound examination is performed prior to the amnioexchange. The patient is admitted to the labor and delivery suite, and a nonstress test is performed before and after the amnioexchange. An intravenous line or a heplock is placed, and a single vial of blood is obtained and held for routine admission laboratory studies, in the case that urgent delivery is required. Prophylactic tocolysis may be given in the form of intravenous or subcutaneous terbutaline. Light IV sedation may also be given if desired.

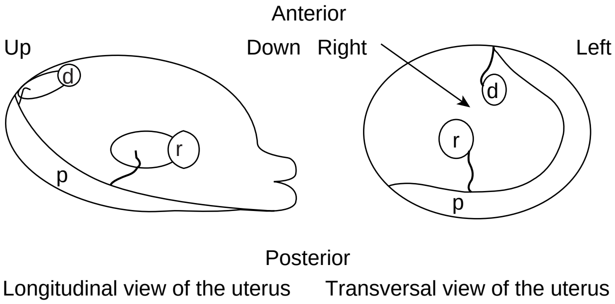

Anterior
Up
Down
Right
Left
d
r
p
c
d
r
p
Posterior
Longitudinal view of the uterus
Transversal view of the uterus

Figure 2. Views of the uterus.

After performing a sterile abdominal prep and drape, amnioexchange is performed using a “closed system.” The closed tubing system materials are illustrated in Fig. 2. All of the materials are sterile except the graduated cylinder. Two sterile three-way stopcocks (b and c) are connected end-to-end. Sterile IV solution (a) is connected to stopcock (c) by way of sterile IV tubing (a). Three lengths of sterile connecting IV tubing (b) are connected to the remaining exposed ports of the stopcock assembly. Tubing from the side port of stopcock (b) is allowed to drain to the graduated cylinder (d). A 60 mL syringe (e) is connected to the tubing attached to the inline port of stopcock (b). The tubing attached to the inline port of stopcock (c) will connect to the therapeutic amniocentesis needle. Stopcock (c) is closed off to the IV solution (a). Stopcock (b) is closed off to the graduate (d). Amniocentesis is performed under continuous ultrasound guidance. Once access is obtained, the stylet is removed from within the needle lumen and the connection tubing is attached. With the stopcocks positioned as noted above, amniotic fluid is withdrawn into the syringe (e) until the syringe is filled. Stopcock (b) is then closed off to the patient and the fluid is expelled into the graduate (d). Stopcock (b) is then turned off to the graduate (d), and this step is repeated until the desired amount of amniotic fluid has been withdrawn (300–900 ml). With stopcock (b) closed to the graduate (d), stopcock (c) is closed off to the patient. The syringe is then filled with sterile warmed saline. Stopcock (c) is then closed to the IV solution (a) and the fluid is infused into the patient. This step is repeated until the desired amount of fluid is infused. These steps are repeated serially until the infusion procedure is complete. If the amniotic fluid volume falls within normal range, the amniotic fluid volume at the end of the amnioexchange should be the same as at the beginning of the procedure. In the presence of oligohydramnios, additional sterile warmed fluid can be added to the uterine cavity in order to achieve a normal fluid volume by the end of the procedure.

If the fluid volume is normal, the placenta is located posteriorly or fundally, and the fetus is quiescent, all of the toxic amniotic fluid planned for removal might be aspirated in one step. With an anteriorly implanted placenta, however, or in the presence of oligohydramnios or an active fetus, it may be necessary to remove a small amount of amniotic fluid, and then replace it with warmed normal saline, repeating the procedure serially until the amnioexchange is completed.

After completion of the amnioexchange, electronic monitoring of the fetal heart rate and uterine activity continues until the patient fulfills the usual criteria for discharge.

AMNIOPATCH

Iatrogenic preterm premature rupture of membranes (PPROM) occurs in approximately 1.2% of patients after genetic amniocentesis (43), 3–5% of patients after diagnostic fetoscopy (44), and approximately 5–8% of patients after operative fetoscopy. Although the membranes might seal spontaneously in this setting (45,46), most patients continue to leak fluid and are at a risk for pregnancy loss.

The overall perinatal mortality of previable PPROM managed expectantly is 60% (47,48). Nearly one-third of these deaths occurs *in utero*. Pulmonary hypoplasia occurs in 50% of cases diagnosed before 19 weeks (49). Serious sequelae in surviving infants include blindness, chronic lung disease, and cerebral palsy.

Patients with iatrogenic PPROM between 16 and 24 weeks gestation who do not have clinical evidence of intra-amniotic infection are candidates for amniopatch therapy. PPROM is confirmed with a sterile speculum examination showing vaginal pooling of fluid, ferning, and a positive Nitrazine test. The maximum vertical pocket of amniotic fluid is measured sonographically. Patients are placed on intravenous antibiotics and bed rest for one week to allow for spontaneous sealing of the membranes. If spontaneous sealing does not occur, 1 unit of autologous platelets and cryoprecipitate are prepared if the patient is eligible for autologous donation. Otherwise, donor platelets and cryoprecipitate are prepared.

After informed consent, an amniocentesis is performed using a 22 gauge needle. The needle is directed into an available pocket of fluid regardless of the site of the previous invasive procedure. A K-51 tubing extension attached to a three-way stopcock is connected to the hub of the needle. Platelets are administered first, followed by cryoprecipitate. In our original protocol, 1 whole unit of platelets was injected. We have subsequently reduced the dose of platelets to one-half a unit because of an unexplained fetal death demise, an adverse effect probably caused by sudden activation of a large number of platelets.

In our series of 28 cases, the average gestational age at the time of the procedure was 19 weeks and 3 days. The average gestational age of delivery in patients who did not have an intrauterine fetal demise was 33 weeks and 4 days. Overall, membrane sealing occurred in 19 of 28 patients (67.9%). Of the 28 patients treated, 11 had a large membrane detachment but no overt leakage of fluid. The detachment of the membrane occurred from fluid escaping the amniotic cavities through the membrane defect, causing dissection of the chorionic cavity. In these patients, only the chorion separates this fluid from leaking grossly to the vagina. In this group, the amniopatch was successful in resealing the amniotic membrane in 7 of the 11 patients (63.6%).

The precise mechanism by which the amniopatch works is unknown. Presumably, platelet activation at the site of rupture and fibrin formation initiates a healing process that enables the membranes to seal.

TTTS

Feto-fetal transfusion syndrome can be described in all monochorionic multiple pregnancies but has been extensively reported in twins. Twin-to-twin transfusion syndrome (TTTS) develops in approximately 15% of all monochorionic pregnancies (50), and carries a high perinatal mortality rate (50). The fetuses are morphologically normal, and inter-twin vascular communications on the chorionic plate are thought to be responsible for the development of the disease through unidirectional blood transfusion from the donor to the recipient twin. Besides the primary hemodynamic imbalance between the twins, the disease may lead to disruptive lesions in both twins. Before the development of antenatal ultrasound, TTTS was diagnosed at birth as a discordance of at least 20% in weight and 5 g/dL in the hemoglobin concentrations of two twins of the same sex (52). These criteria were abandoned because these features could not be consistently recognized in utero. With the development of ultrasound, the polyhydramnios-oligohydramnios sequence has been found to be the condition carrying one of the highest perinatal mortality rates in obstetrics, up to 90% without treatment.

Laser coagulation of placental anastomoses by fetoscopy is the most effective first-line treatment of FFTS, which leads to at least one survivor at birth and intact survival at 6 months of age in 76% and 76% respectively, as compared with 56% and 51% in cases treated by serial amnioreduction in the Eurofetus randomized trial (53).

The selection criteria to qualify for percutaneous endoscopy-directed laser coagulation of placental anastomoses include:

1. Gestational age of less than 26 weeks.
2. Ultrasound diagnosis of a single monochorionic placenta by ultrasound in the first trimester of pregnancy.
3. Polyhydramnios in the recipient's amniotic cavity with a deepest vertical pool $\geq$8 cm or $\geq$10 cm before or after 20 weeks of gestation, respectively.
4. Oligohydramnios in the donor's amniotic sac with a deepest vertical pool $\leq$2 cm.

Preoperative evaluation consists of ultrasound examination, including morphological examination, fetal Doppler, cardiothoracic index, identification of placental location, and cord insertions. Amniocentesis or amnioreduction prior to laser may cause intra-amniotic bleeding and therefore make the procedure more difficult, or even impossible, due to impaired visualization. The site of entry is chosen as demonstrated in Fig. 1, for the scope to be entered at a right angle to the long axis of the small twin in order to maximize the chance to ensure adequate visualization of the placental surface and intertwin membranes. Ideally, the scope should also be entered alongside a virtual line joining the two cord insertions. When these criteria are met, the vascular equator of the placenta as well as the vascular anastomoses on the chorionic plate are more likely to be visualized in the operative field.

Prophylactic cefazolin 2 g, indomethacin suppository 100 mg, and oral flunitrazepam are given before surgery and local anesthesia with nonadrenalinized xylocaine is injected down to the myometrium. A 10 Fr cannula for a central venous catheter loaded with a trocar is introduced percutaneously under continuous ultrasound guidance. A 2 mm 0° fetoscope (Storz 26008 AA) is passed down a straight or curved sheath to operate on posterior or anterior placentas, respectively. The sheath also has a working channel carrying a 1 mm diode laser fiber.

A systematic examination of the chorionic plate alongside the insertion of the inter-twin membrane is performed. Identification of crossing vessels and of their arterial or venous nature is possible because arteries cross over veins and show a darker red color than veins owing to a lower oxygen saturation in the circulating blood (54). Selective coagulation of anastomotic vessels is performed with the aim of separating the monochorionic placenta into two distinct fetal-placental circulations, sparing the normal cotyledons of each placental territory. Nonselective coagulation of crossing vessels is only performed when the distal end or the origin of the vessel cannot be identified. The power of the diode laser is set at 30–50 w. At the end of theprocedure, excessive amniotic fluid is drained through the sheath of the fetoscope until normal amniotic fluid volume is obtained with a deepest vertical pool of 5–6 cm.

BIPOLAR UMBILICAL CORD OCCLUSION

Selective reduction in complicated monochorionic (MC) multifetal pregnancies is performed to prevent the delivery of an anomalous or severely compromised fetus and improve the perinatal outcome for the surviving co-twin by delaying delivery or risk associated with spontaneous loss of the affected. The use of cardiotoxic agents, such as potassium chloride, is contraindicated in MC pregnancies because of the potential vascular transmission of the agent and compromise of the co-twin due to the presence of placental vascular anastomoses. Thermal vascular occlusive techniques, such as bipolar umbilical cord occlusion (BPC), have been shown to achieve the stated goals with minimal maternal morbidity. Indications for BPC that are unique to MC gestations include twin reverse arterial perfusion, discordant fetal anomalies, and isolated severe growth lag. BPC has also been used as a primary intervention in advanced twin-twin transfusion syndrome or as a secondary procedure when alternative therapies such as amnioreduction or laser have failed to correct the disease process.

The procedure was originally described by Deprest et al. (55). In brief, using the standard sterile technique, the patient's abdomen is properly prepared and draped. An abdominal ultrasound is performed to confirm fetal position, viability, and umbilical cord locations. General and conduction anesthesia may be used; however, intravenous sedation with local infiltration of 1% lidocaine or 0.25% bupivicane for subcutaneous, deep muscle, and fascia anesthesia is usually sufficient and is associated with less maternal morbidity. A small skin incision is made to allow insertion of an endoscopic trocar. Under continuous ultrasound guidance, the instrument is inserted through a placental free window toward the targeted umbilical cord, ideally avoiding the gestational sac of the normal co-twin. Once the trocar is secured in the amniotic sac, the obturator is removed. The bipolar forceps are inserted and advanced to the umbilical cord.

The cord is grasped and positioned away from the amnion before thermal energy is applied. The duration and wattage (W) necessary for occlusion will vary, from 20–60 s and 20–50 W, respectively, based on the gestational age and umbilical cord thickness. When a full-thicknessgrasp exists, application of the thermal energy will result in turbulence and "streaming" of amniotic fluid adjacent to the forceps. It is not uncommon to have an audible "pop" secondary to the heating of Wharton's jelly and subsequent rupture of amnion at the site of occlusion, which should not be perceived as a sign of completed coagulation. As a result of the natural spiral of the umbilical cord, complete occlusion of all vessels requires 2–3 applications of the forceps at adjacent sites. Pulse and color flow Doppler blood flow studies are performed to confirm cord occlusion at each site.

The size of the BPC forceps that have been used for these procedures has varied from 2.2–5.0 mm. The majority of procedures have been performed with commercially available single-use 3.0 mm bipolar diathermy forceps (55–58).

Intravenous prophylactic antibiotics and indomethacin for tocolysis are generally given prior to the procedure. Postoperative monitoring for uterine contractions and, depending on the gestational age, continuous or intermittent fetal heart rate should be done for at least 2 hours. The majority of programs will observe patients for an extended period of 12–24 h with limited activity. Subsequent doses of antibiotics and tocolytic treatment are given during this time. Prior to discharge, a limited ultrasound is performed to determine the amniotic fluid volume and assess for signs of hydrops and anemia, including Doppler velocemitry of the middle cerebral artery and, where appropriate, similar studies of the umbilical artery and ductus venosus. If no evidence of preterm labor, leaking of amniotic fluid, or bleeding exists, the patient is discharged with instructions to continue with modified bed rest at home for 7–10 days, take her temperature bid, and report an elevation, leaking of vaginal fluid, bleeding, or contractions. An ultrasound is performed in 10–14 days and then at a minimum every 4 weeks thereafter. Additional ultrasounds and fetal monitoring should be performed as clinically indicated by the primary disease and gestational age.

BIBLIOGRAPHY

1. Harrison MR, Adzick NS. Open Fetal Surgery Techniques. The Unborn Patient: The Art and Science of Fetal Therapy, 3rd ed. Harrison MR, Evan MI, Adzick NS, Holzgreve W, editors. New York: WB Saunders Company; 2001. pp 247–255.
2. Harrison MR, Anderson J, Rosen MA, et al. Fetal surgery in the primate I. Anesthetic, surgical and tocolytic management to maximize fetal-neonateal survival. J Pediatr Surg 1982;17:115–122.
3. Nakayama DK, Harrison MR, Seron-Ferre M, et al. Fetal surgery in the primate II. Uterine electromyographic response to operative procedure and pharmacologic agents. J Pediatr Surg 1984;19:333–339.
4. Adzick NS, Harrison MR, Glick PL, et al. Fetal surgery in the primate III. Maternal outcome after fetal surgery. J Pediatr Surg 1986;21:477–480.
5. Bianchi DW, Crombleholme TM, D'Alton ME. Cystic Adenomatoid Malformation. In: Bianchi DW, Crombleholme TM, D'Alton ME, editors. Fetology-Diagnosis & Management of the Fetal Patient. New York: McGraw-Hill; 2000. pp 289–297.
6. Bianchi DW, Crombleholme TM, D'Alton ME. Sacrococcygeal teratoma. In: Bianchi DW, Crombleholme TM, D'Alton ME, editors. Fetology-Diagnosis & Management of the Fetal Patient. New York: McGraw-Hill; 2000. pp 867–877.
7. Altman RP, Randolph JG, Lilly JR. Sacrococcygeal teratoma: American Academy of Pediatrics Surgical Section Survey 1973. J Pediatr Surg 1974;9:389–398.
8. Adzick NS. Management of fetal lung lesions. Clin Perinatol 2003;30:481–492.
9. Hedrick HL, Flake AW, Crombleholme TM, et al. Sacrococcygeal teratoma: Prenatal assessment, fetal intervention, and outcome. J Pediatr Surg 2004;39(3):430–438.
10. Crombleholme TM, Coleman B, Hedrick H, et al. Cystic adenomatoid malformation volume ratio predicts outcome in prenatally diagnosed cystic adenomatoid malformation of the lung. J Pediatr Surg 2002;27(3):331–338.
11. Rice HE, Estes JM, Hedrick MH, et al. Congenital cystic adenomatoid malformations: A sheep model. J Pediatr Surg 1994;29:692–696.
12. Flake AW. Fetal sacrococcygeal teratoma. Sem Pediatr Surg 1993;2:113–120.
13. Adzick NS, Crombleholme TM, Morgan MA, et al. A case report. A rapidly growing fetal teratoma. Lancet 1997;349:538.
14. Hedrick HL. Ex utero intrapartum therapy. Semi Ped Surg 2003;10(3):190–195.
15. Wilson RD, Johnson MP, Flake AW, et al. Reproductive outcomes after pregnancy complicated by maternal-fetal surgery. Am J Obstet Gynecol 2004;191:1430–1436.
16. Harrison MR, Filly RA, Parer JT, et al. Management of the fetus with a urinary tract malformation. JAMA 1981; 246(6):635–639.
17. Nakayama D, Harrison M, deLorimier A. Prognosis of posterior urethral valves present at birth. J Ped Surg 1986;21:43–45.
18. Golbus MS, Harrison MR, Filly RA, et al. In utero treatment of urinary tract obstruction. Am J Obstet Gynecol 1982; 383–388.
19. Berkowitz RL, Glickman MG, Smith GJ, et al. Fetal urinary tract obstruction: What is the role of surgical intervention in utero? Am J Obstet Gynecol 1982;144(4):367–375.
20. Rodeck C, Nicolaides K. Ultrasound guided invasive procedures in obstetrics. Clin Obstet Gynecol 1983;10:515.
21. Beck AD. The effect of intra-uterine urinary obstruction upon the development of the fetal kidney. Urol 1971;105:784–789.
22. Pringle KC, Bonsib SM. Development of fetal lamb lung and kidney in obstructive uropathy: A preliminary report. Fetal Ther 1988;3(1-2):118–128.
23. Manning FA, Harman CR, Lange IR, et al. Antepartum chronic fetal vesicoamniotic shunts for obstructive uropathy: A report of two cases. Am J Obstet Gynecol 1983;145(7):819–822.
24. Young H, Frontz W, Baldwin J. Congenital obstruction of the posterior urethra. J Urol 1919;3:289–365.
25. Reuss A, Wladimiroff J, Niermeyer M. Antenatal diagnosis of renal tract anomalites by ultrasound. Pediat Nephrol 1987;1:546–552.
26. Johnson MP, Bukowski TP, Reitleman C, et al. In utero surgical treatment of fetal obstructive uropathy: A new comprehensive approach to identify appropriate candidates for vesicoamniotic shunt therapy. Am J Obstet Gynecol 1994;170(6):1770–1776; discussion 1776–1779.
27. Quintero RA, Hume R, Smith C, et al. Percutaneous fetal cystoscopy and endoscopic fulguration of posterior urethral valves [see comments]. Am J Obstet Gynecol 1995;172(1 Pt 1):206–209.
28. Quintero RA, Johnson MP, Romero R, et al. In-utero percutaneous cystoscopy in the management of fetal lower obstructive uropathy. Lancet 1995;346(8974):537–540.
29. Quintero RA, Shukla AR, Homsy YL, et al. Successful in utero endoscopic ablation of posterior urethral valves: A new dimension in fetal urology. Urology (Online) 2000;55(5): 774.
30. Steinbok P, Irvine B, Cochrane DD, Irwin BJ. Long-term outcome and complications of children born with myelomeningocele. Child's Nerv Syst 1992;8:92–96.
31. McLone DG. Continuing concepts in the management of spina bifida. Pediare Neurosurg 1992;18:254–257.
32. Tulipan N, Hernanz-Schulman M, Bruner JP. Reduced hindbrain herniation after intrauterine myelomeningocele repair: A report of four cases. Pediatr Neurosurg 1998;29: 274–278.
33. Tulipan N, Hernanz-Schulman M, Bruner JP. Intrauterine myelomeningocele repair reverses preexisting hindbrain herniation. Pediatr Neurosurg 1999;31:137–142.
34. Bruner JP, Tulipan N, Paschall RL, Boehm FH, Walsh WF, Silva SR, Hernanz-Schulman M, Lowe LH, Reed GW. Fetal surgery for myelomeningocele and the incidence of shunt-dependent hydrocephalus. JAMA 1999;282:1819–1825.
35. Johnson MP, Sutton LN, Rintoul N, Crombleholme TM, Flake AW, Howell LJ, Hedrick HL, Wilson RD, Adzick NS. Fetal myelomeningocele repair: Short-term clinical outcomes. Am J Obstet Gynecol 2003;189:482–487.
36. Bruner JP, Boehm FH, Tulipan N. The Tulipan-Bruner trocar for uterine entry during fetal surgery. Am J Obstet Gynecol 1999;181:1188–1191.
37. Mangels KJ, Tulipan N, Bruner JP, Nickolaus D. Use of bipedicular advancement flaps for intrauterine closure of myeloschisis: Technical report. Pediatr Neurosurg 2000; 32:52–56.
38. Kohl T, Sharland G, Allan LD, Gembruch U, Chaoui R, Lopes LM, Zielinsky P, Huhta J, Silverman NH. World experience of percutaneous ultrasound-guided balloon valvuloplasty in human fetuses with severe aortic valve obstruction. Am J Cardiol 2000;85:1230–1233.
39. Tworetzky W, Wilkins-Haug L, Jennings RW, van der Velde ME, Marshall AC, Marx GR, Colan SD, Benson CB, Lock JE, Perry SB. Balloon dilation of severe aortic stenosis in the fetus: Potential for prevention of hypoplastic left heart syndrome: candidate selection, technique, and results of successful intervention. Circulation 2004;110:2125–2131.

40. Luton D, de Lagausie P, Guibourdenche J, Peuchmaur M, Sibony O, Aigrain Y, Oury IF, Blot P. Influence of amnioinfusion in a model of in utero created gastroschisis in the pregnant ewe. Fetal Diagn Ther 2000;15:224–228.
41. Luton D. Etude de L'inflammation, dans le Laparoschisis, Humain et dans un Modele de Laparoschisis de Bredis. These; 2001.
42. Luton D, de Lagausie P, Guibourdenche J, Oury IF, Sibony O, Vuillard E, Boissinot C, Aigrain Y, Beaufils F, Navarro J, Blot P. Effect of amnioinfusion on the outcome of prenatally diagnosed gastroschisis. Fetal Diagn Ther 1999;14:152.
43. The NICHD National Registry for Amniocentesis Study Group. Midtrimester amniocentesis for prenatal diagnosis. Safety and accuracy. JAMA 1976;236:1471–1476.
44. Rodeck C. Fetoscopy guided by real-time ultrasound for pure fetal blood samples, fetal skin samples, and examination of the fetus in utero. Br J Ob Gyn 1980;87:449–456.
45. Quintero R, Reich H, Puder K, et al. Brief report: Umbilical-cord ligation of an Acardiac twin by fetoscopy at 19 weeks of gestation. N Engl J Med 1994;33:469–471.
46. Gold R, Goyert G, Schwartz D. Conservative management of second-trimester post-amniocentesis fluid leakage. Obstet Gynecol 1989;74:745–747.
47. Bengtson J, VanMarter L, Barss V. Pregnancy outcome after premature rupture of the membranes at or before 26 weeks' gestation. Obstet Gynecol 1989;73:921–927.
48. Beydoun S, Yasin S. Premature rupture of the membranes before 28 weeks: Conservative management. Am J Obstet Gynecol 1986;155:471–479.
49. Rotschild A, Ling E, Puterman M. Neonatal outcome after prolonged preterm rupture of the membranes. Am J Obstet Gynecol 1990;162:46–52.
50. Sebire NJ, Snijders RJ, Hughes K, et al. The hidden mortality of monochorionic twin pregnancies. Br J Obstet Gynecol 1997;104(10):1203–1207.
51. Patten RM, et al. Disparity of amniotic fluid volume and fetal size: Problem of the stuck twin-US studies. Radiology 1989;172:153–157.
52. Danskin FH, Neilson JP. Twin-to-Twin transfusion syndrome: What are appropriate diagnostic criteria? Am J Obstet Gynecol 1989;161:365–369.
53. Senat MV, Deprest J, Boulvain M, Paupe A, Winer N, Ville Y. Endoscopic laser surgery versus serial amnioreduction for severe twin-to-twin transfusion syndrome. N Engl J Med 2004;351:136–144.
54. Benirschke K, Driscoll S. The Pathology of the Human Placenta. New York: Springer-Verlag; 1967.
55. Deprest JA, Audibert F, Van Schoubroeck D, Hecher K, Mahieu-Caputo D. Bipolar coagulation of the umbilical cord in complicated monochorionic twin pregnancy. Am J Obstet Gynecol 2000;182:340–345.
56. Nicolini U, Poblete A, Boschetto C, Bonati F, Roberts A. Complicated monochorionic twin pregnancies: Experience with bipolar cord coagulation. Am J Obstet Gynecol 2001;185:703–707.
57. Taylor MJO, Shalev E, Tanawattanacharoen S, Jolly M, Kumar S, Weiner E, Cox PM, Fisk NM. Ultrasound-guided umbilical cord occlusion using bipolar diathermy for Stage III/IV twin-twin transfusion syndrome. Prenat Diagn 2002; 22:70–76.
58. Johnson MP, Crombleholme TM, Hedrick H, et al. Bipolar umbilical cord cauterization for selective termination of complicated monochorionic pregnancies. Am J Obstet Gynecol 2001;185(6):S245.

See also FETAL MONITORING; HYDROCEPHALUS, TOOLS FOR DIAGNOSIS AND TREATMENT OF; MONITORING, UMBILICAL ARTERY AND VEIN.

OIL. See LENSES, INTRAOCULAR.

ION-EXCHANGE CHROMATOGRAPHY. See CHROMATOGRAPHY.

IONIZING RADIATION, BIOLOGICAL EFFECTS OF

ZELENNA GOLDBERG
Department of Radiation
Oncology
Davis, California

INTRODUCTION

Radiation biology refers to all biologic responses induced in cells and tissues by ionizing radiation, a term that encompasses the study of all action of ionizing radiation on living things. Ionizing radiation (IR) is radiation that has sufficient energy to cause the ejection of an orbital electron from its path around the nucleus, which indicates that the photon or charged particle can release large amounts of energy in a very small space. IR is separated into electromagnetic radiation and particulate radiation. Electromagnetic radiation consists of X rays or γ rays, which differ only in their mode of production, not in their physical effects in interacting with biologic tissue. X rays are produced by machines that accelerate electrons and then focus them to hit a target usually made of tungsten or gold. The kinetic energy is transferred from the electrons to the target and then is released as photons. These are the most common medical exposures through diagnostic or therapeutic radiation. In contrast, γ-radiation is produced within the nucleus from radioactive isotopes. The γ rays result from the unstable nuclear configuration decaying to a more stable state and releasing the "extra" energy in this transition in the form of the γ ray. Natural background radiation is of this type. Particulate radiation is those radiation sources that are not photons and consist of electrons, protons, α-particles, neutrons, and heavy charged particles such as the nuclei of carbon, neon, argon, or iron. These particles are positively charged because the orbiting electrons have been stripped from them.

As noted, IR is defined by its causing the ejection of an orbital electron when the radiation interacts with tissue. These ionizing events can be further subdivided by being directly ionizing or indirectly ionizing. Directly ionizing events are those in which a charged particle with sufficient kinetic energy interacts with the tissue, directly resulting in the ejection of the orbital electron. These events happen frequently within an exposed tissue, so the charged particle rapidly transfers its energy to the tissue, a concept codified as linear energy transfer (LET). Charged particles have a high LET. In contrast, indirectly ionizing radiation such as γ- or X-radiation is first absorbed in the material and secondary, energetic charged particles (electrons) are released. Because this latter process only occurs when the photon is close enough to an atom to interact, it is referred to as sparsely ionizing radiation and has a low linear energy transfer (LET).

A longstanding paradigm in radiation biology has been that many effects induced by IR, including its carcinogenic effects and ability to kill cancer cells, are the result of DNA damage arising from the actions of IR in cell nuclei, especially interactions of IR and its products with nuclear DNA (1–3). When a charged particle or secondary electron produced by the interaction of a photon with an orbital electron damages DNA itself, it is known as the direct action of IR. Yet, DNA represents a small fraction of the actual size of the cell. Therefore, it was recognized that most of the interaction of IR within a cell would be with more abundant biomolecules, specifically water. If the action is mediated through the intracellular production of radiolytic reactive products, e.g., $OH^{\cdot}$, $H^{\cdot}$, $O_2^{\cdot}$, and H_2O_2, that are generated in aqueous fluid surrounding DNA, then the DNA damage is called indirect action. These processes unquestionably can result in a variety of types of DNA damage, including DNA single- and double-strand breaks, modifications of deoxyribose rings and bases, intra- and inter-strand DNA–DNA cross-links, and DNA–protein cross-links (1,4,5). About a third of all DNA damage is caused by the direct effects of sparsely ionizing γ and X rays, with the remaining balance being attributable to the indirect actions of IR. With high-LET radiation such as the more densely ionizing α-particles that are emitted by radon and radon progeny, the direct actions of IR on DNA become more predominant and the nature of the DNA modifications become much more complex. Regardless of the type of IR, all of the above forms of DNA damage can lead to untoward effects in cells if unrepaired or misrepaired. With specific regard to carcinogenesis, genomic mutations caused by IR are widely thought to arise from DNA damage that is subsequently converted into a mutation as a result of misprocessing by DNA repair mechanisms or that is converted into a heritable mutation when DNA undergoes replication.

This classic view of radiation biology is giving way to a more complex and complete understanding that the cell membrane and other intracellular compartments, as well as the tissue vasculature, are important targets of IR. Furthermore, as detailed below, although intracellular effects are better defined, we now recognize that the extracellular environment and cell–cell contact play important roles in modulating the effects of IR at the cellular and tissue levels.

MOLECULAR RADIATION BIOLOGY/BYSTANDER EFFECTS

IR is a cellular toxin that is sensed at the cellular level through the ATM-p53-p21 pathway (6,7). Although the upstream sensor of ATM remains to be definitively elucidated, the initial radiation sensing protein likely has a redox sensor and undergoes a conformational change, or possibly is phosphorylated, resulting in the phosphorylation of ATM (8). This, in turn, activates the p53 pathway leading to cell cycle arrest, nuclear translocation of transcription factors such as NF-κB, and either cellular repair or apoptosis. How an individual cell commits to a given fate (repair or apoptosis) remains unclear, but undoubtedly it represents the final integrated response to many simultaneous intracellular events. Several excellent reviews on this topic have been written (9,10). It should be noted that IR also affects the 26s proteosome, which is another level of non-nuclear intracellular response affecting cell survival, presumably through alterations in the removal of activated (phosphorlyated) proteins, or proteins active in apoptotic processes such as Bcl2 (11,12).

Radiation effects in cells not directly hit by a radiation ionizing event are called bystander effects. Although initial interest in this effect can be found in the medical literature as far back as into the 1950s, the current interest in the area was stimulated by a study published in 1992 in which Nagasawa and Little (13) observed increases in the frequency of sister chromatid exchanges in ~30% of immortalized Chinese hamster ovary cells that received low-dose exposure to α-particles, even though less than 1% of the cells' nuclei were estimated to actually receive direct nuclear hits by an α-particle. That a relatively low percentage of the cells experienced one or more direct "hits" by the α-particles, be they in the cytoplasm or in the nuclei, suggested the possibility that some mechanism was conveying a radiation-associated response to unirradiated cells. Other groups went on to confirm and extend on this finding. It is now well recognized that cells do not require a direct nuclear traversal to result in radiation changes. There are extracellular responses, predominantly TGF-β mediated through media (cell culture experiments) or extracellular fluids (tissues), but other factors cannot be excluded (14). Furthermore, it is recognized that cell–cell contact and gap junctional communications are critical in transmitting the signal from the directly irradiated cell to the neighboring, bystander cells (15).

TIME, DOSE, AND FRACTIONATION

The biologic effects of IR in tissue relate to the size of the dose delivered, time between radiation exposures and the total dose of IR given. Although environmental exposures are chronic and (usually) low level, medical exposures are acute and can be repetitive when given for the treatment of cancer. Radiation therapy for the treatment of malignancy remains the most effective anti-cancer agent discovered, and treatment schedules are predicated on the "4 R's of radiobiology": repair, repopulation, redistribution, and reoxygenation.

Repair of radiation-induced damage underlies the intrinsic radiation sensitivity of the cell to radiation cell killing. Cells can be broadly grouped into those that can repair significant amounts of damage and those with more limited repair capacity. The former descriptive category corresponds to tissue types where there is limited normal cell turnover, such as lung, kidney, or brain, and these are tissues that display damage after a more prolonged time and are therefore known as late responding tissues. The cells with more limited intrinsic repair capacity are known as acute responding tissues and are typified by skin or gut. These cell types have a limited life span and are constantly being replaced within normal physiologic functioning.

Repopulation refers to the generation of new cells to replace those killed by the IR exposure. Although therapeutically beneficial for containing normal tissue toxicity, repopulation is also active in malignant tissue and thereby allows greater numbers of tumor clonogens to develop after treatment. For some types of tumors, an acceleration of repopulation begins after 4 weeks of radiation therapy, which can compromise clinical outcome if prolonged therapeutic IR fractionation schemes are used.

Redistribution refers to the changes in cell assortment across the cell cycle. It has long been established that cells vary in their sensitivity to the cytotoxic effects of IR depending on where in the cell cycle they are at the time of irradiation (16,17). Cells are most sensitive to IR effects in the G2/M phase and are most resistant during the S-phase. G1 is intermediate between these two. Thus, clinical radiation therapy schedules use multiple fractions to overcome the relative resistance of the cells in S-phase of the cell cycle on a given treatment day. This difference in radioresistance across stages in the cell cycle also underlies one mechanism of the synergy between radiation therapy and many chemotherapeutic agents in the treatment of malignant disease. Although S-phase cells resist killing from the IR, they are more sensitive to some chemotherapeutics that are active during S-phase when the DNA is most exposed and is replicating. Furthermore, this alteration of cell cycle sensitivity to IR cytotoxicity also has been an area of research for IR-biologic response modifiers. If one can increase the percentage of cells in G2/M at the time of IR, then the relative cell kill per fraction of IR will increase.

Reoxygenation refers to the presence or absence of molecular oxygen within the cell at the time of IR delivery. As detailed at the beginning, most of the cellular damage from IR is mediated through the production of free radicals within the cell. If molecular oxygen is present, these free radicals can be transformed into more complex peroxide radicals, which are more difficult for the cell to repair. This is classically known as "fixing the radiation damage" in the British use of the term "fix" (make more solid), not the American (repair). The increase in cell killing from IR in the presence of oxygen versus under hypoxic conditions is known as the oxygen enhancement ratio, and it is approximately a factor of 2–3, dependent on where in the cell cycle the irradiated cell sits. Although normal tissues have a well-maintained vascular supply so that oxygenation is essentially constant, tumors have a tortuous and unstable vasculature, so that the vessels can open and close erratically. This type of hypoxia is referred to as "acute hypoxia." Chronic hypoxia occurs when the tumor outgrows its blood supply and the tumor cells are simply beyond the diffusion capability of molecular oxygen. Attempts to exploit this differential, either by sensitizing the hypoxic cells or by specifically targeting them, have been explored and remain active areas of research. Furthermore, identification of the presence of hypoxia remains a significant clinical strategy (18).

The four R's of radiobiology reflect intracellular controls of overall radiation response. The tissue level effects from IR in the treatment of cancer are dependent on several parameters: volume, dose per fraction, total dose, and time between fractions. The volume of tissue irradiated is critical for determining the long-term repair by repopulation by normal cells to fill in for those irreparably damaged by the IR. Each cell type has its own intrinsic radiation sensitivity, and the tissue organization (serial or parallel cellular arrangement) as well as the tissue vasculature play critical roles in tissue repair and thus radiosensitivity. Tissues are subdivided into two categories of radiosensitivity based on when they display their damage. Tissues that display their radiation induced damage during a standard course of medical radiation therapy are known as "acute responding tissues." These tissues naturally have a large amount of cellular turnover, such as skin or gut, and at a cellular level, this corresponds to lesser ability to repair sublethal DNA damage (i.e., a larger proportion of DNA damage is lethal and nonrepairable). In contrast, tissues where there is little or no normal cellular turnover, such as brain, kidney, lung, or spinal cord, there is a substantial ability to repair sublethal damage, and tissue toxicity from radiation therapy is displayed late, long after therapy has finished. These tissues are therefore labeled "late responding tissues." Therapeutic strategies are designed to separate these two types of responses as malignant tumors are models of acute responding tissues, whereas the dose limiting side effects from radiation therapy are secondary to late responding tissue effects (19,20).

THE LINEAR NO-THRESHOLD MODEL OF RADIATION EFFECTS

Based on the data collected from the victims of the bomb detonations at Hiroshima and Nagasaki, a linear no-threshold model of radiation effects was developed and adopted for public policy applications. In essence, the model states that (1) all radiation exposures are biologically active, (2) the response in the cells/tissue is linear with dose, and (3) there is no threshold below which there is no or negligible effects. The model was developed from the moderately low-dose exposure ranges of 0.5–2 Gy and the medically significant sequelae of increased mortality (carcinogenic and noncarcinogenic, predominantly cardiovascular) (21–23). Hiroshima and Nagasaki data represent the effects of a single acute dose exposure delivered to the whole body with the nutritional deprivation from WW2; as such, they are subject to criticism and questions of their applicability to modern healthy populations where low-dose radiation exposure is often of a more chronic nature. Nevertheless, they remain the best available population-based data, and they have been painstakingly collected and analyzed. As detailed below, however, the shape of the response curve at the lowest doses remains controversial.

LOW-DOSE EXPOSURES

Low-dose ionizing radiation (LDIR) in the 1–10 cGy range has largely unknown biological activity in the human. Current modeling for health and safety regulations, as well as prediction of carcinogenesis, presupposes a linear, no-threshold model of radiation effects based on the nuclear bomb explosion data, which estimates the effect and risk at low dose by extrapolation from measured effects at high doses. Yet the scientific literature presents a more complex picture, and few data clearly support a linear dose-response model and none in humans. Numerous studies suggest some effects of LDIR may be benign or even beneficial under some circumstances (24,25). Reported benefits include stimulated growth rates in animals, increased rates of wound healing, reductions in cell apoptosis, enhancements in the repair of damaged DNA, and increases in radioresistance via the induction of an adaptive response, among others (26–28). Other lines of evidence, however, suggest LDIR can be hazardous, and if a threshold for detrimental responses does exist, it is operational only at very low dose levels, e.g., $\leq$1 cGy. Schiestl et al. (29) found that 1–100 cGy doses of X rays could cause genetic deletions in mice in a linear dose-response manner. This remains an active area of research (30).

Little is known regarding individual variability in sensitivity to radiation exposure. Studies are actively ongoing to develop methods to best assess interindividual variability. This understanding will have both a health risk assessment and medical applications (31).

GENOMIC INSTABILITY

Radiation-induced genomic instability encompasses a range of measurable endpoints such as chromosome destabilization, sister chromatid exchanges, gene mutation and amplification, late cell death, and aneuploidy, all of which may be causative factors in the development of clinical disease, including carcinoma.

Kadhim et. al. identified the persistence of radiation-induced chromosomal instability following α-particle irradiation in clonal populations of murine bone marrow cells (32). The same group followed up this seminal work with an examination of four human bone marrow samples, subjected to an *ex vivo* α-particle IR (33). Further research then demonstrated that the genomic instability phenotype could be transmitted *in vivo* when murine hemopoietic cells that had been irradiated *in vitro* were transplanted into mice that had previously had their native bone marrow purged (34). However, when Whitehouse and Tawn examined radiation workers in Sellafield, England, who had bone marrow plutonium deposition evidence for genomic instability was not found (35). Whether γ-IR could induce genomic instability was then examined. The original reports from Kadhim et al. were negative (32,33), but further research examining hprt locus mutations convincingly demonstrated that genomic instability was inducible by γ-irradiation (36). Thus, although genomic instability can clearly be demonstrated in the laboratory, whether it occurs in humans after IR exposure remains uncertain (37).

Our understanding of the biologic effects of IR is evolving and ever growing. Research has been active in this field for over 100 years, and it remains a vibrant research area with the new molecular tools now available. The target of interest and concern is as small as the individual DNA base or as large as the whole organism. Mechanistic studies of the subcellular targets of IR and the cellular response signaling cascades must be matched with more complex system evaluations so that the summative effect of these pathways becomes known. Cell signaling exists within and between cells, and this cross-talk affects the ultimate response to IR exposure. Improving our understanding of radiation response at the cellular and tissue levels will undoubtedly yield advances for medical/therapeutic radiation as well as general cellular biology.

BIBLIOGRAPHY

1. Goodhead DT. Initial events in the cellular effects of ionizing radiations: Clustered damage in DNA. Int J Radiation Biol 1994;65(1):7–17.
2. Iliakis G. The role of DNA double strand breaks in ionizing radiation-induced killing of eukaryotic cells. Bioessays 1991;13(12):641–648.
3. Sutherland BM, et al. Clustered DNA damages induced by x rays in human cells. Radiat Res 2002;157(6):611–616.
4. Ward JF. DNA damage produced by ionizing radiation in mammalian cells: identities, mechanisms of formation, and reparability. Prog Nucleic Acid Res Mol Biol 1988;35:95–125.
5. Sutherland BM, et al. Clustered DNA damages induced in isolated DNA and in human cells by low doses of ionizing radiation. Proc Natl Acad Sci USA 2000;97(1):103–108.
6. Fernandes N, et al. DNA damage-induced association of ATM with its target proteins requires a protein interaction domain in the N terminus of ATM. J Biol Chem 2005;280(15):15158–15164.
7. Kang J, et al. Functional interaction of H2AX, NBS1, and p53 in ATM-dependent DNA damage responses and tumor suppression. Mol Cell Biol 2005;25(2):661–670.
8. Lavin MF, et al. ATM signaling and genomic stability in response to DNA damage. Mutat Res 2005;569(1–2):123–132.
9. McBride WH, et al. A sense of danger from radiation. Radiat Res 2004;162(1):1–19.
10. Li L, Zou L. Sensing, signaling, and responding to DNA damage: Organization of the checkpoint pathways in mammalian cells. J Cell Biochem 2005;94(2):298–306.
11. Ghobrial IM, Witzig TE, Adjei AA. Targeting apoptosis pathways in cancer therapy. CA Cancer J Clin 2005;55(3):178–194.
12. Pervan M, et al. Molecular pathways that modify tumor radiation response. Am J Clin Oncol 2001;24(5):481–485.
13. Nagasawa H, Little JB. Induction of sister chromatid exchanges by extremely low doses of alpha- particles. Cancer Res 1992;52(22):6394–6396.
14. Barcellos-Hoff MH. Integrative radiation carcinogenesis: Interactions between cell and tissue responses to DNA damage. Semin Cancer Biol 2005;15(2):138–148.
15. Goldberg Z, Lehnert BE. Radiation-induced effects in unirradiated cells: A review and implications in cancer. Int J Oncol 2002;21:337–349.
16. Brown JM. The effect of acute x-irradiation on the cell proliferation kinetics of induced carcinomas and their normal counterpart. Radiat Res 1970;43(3):627–653.
17. Brown JM, Berry RJ. Effects of X-irradiation on the cell population kinetics in a model tumour and normal tissue system: Implications for the treatment of human malignancies. Br J Radiol 1969;42(497):372–377.
18. Brown JM, Wilson WR. Exploiting tumour hypoxia in cancer treatment. Nat Rev Cancer 2004;4(6):437–447.
19. Fowler JF. The eighteenth Douglas Lea lecture. 40 years of radiobiology: Its impact on radiotherapy. Phys Med Biol 1984;29(2):97–113.
20. Fowler JF. Potential for increasing the differential response between tumors and normal tissues: Can proliferation rate be used? Int J Radiat Oncol Biol Phys 1986;12(4):641– 645.
21. Hayashi T, et al. Long-term effects of radiation dose on inflammatory markers in atomic bomb survivors. Am J Med 2005;118(1):83–86.

22. Preston DL, et al. Effect of recent changes in atomic bomb survivor dosimetry on cancer mortality risk estimates. Radiat Res 2004;162(4):377–389.
23. Yamada M, et al. Noncancer disease incidence in atomic bomb survivors, 1958–1998. Radiat Res 2004;161(6):622–632.
24. Loken MK, Feinendegen LE. Radiation hormesis. Its emerging significance in medical practice. Invest Radiol 1993;28(5):446–450.
25. Jaworowski Z. Hormesis: The beneficial effects of radiation. 21st Century Sci Tech 1994;7:22–27.
26. Shadley JD, Afzal V, Wolff S. Characterization of the adaptive response to ionizing radiation induced by low doses of X rays to human lymphocytes. Radiat Res 1987;111(3):511–517.
27. Liu SZ, Liu WH, Sun JB. Radiation hormesis: Its expression in the immune system. Health Phys 1987;52(5):579–583.
28. Sagan LA, Cohen JJ. Biological effects of low-dose radiation: Overview and perspective. Health Phys 1990;59(1): 11–13.
29. Schiestl RH, Khogali F, Carls N. Reversion of the mouse pink-eyed unstable mutation induced by low doses of x-rays. Science 1994;266(5190):1573–1576.
30. Goldberg Z, et al. Effects of low-dose ionizing radiation on gene expression in human skin biopsies. Int J Radiat Oncol Biol Phys 2004;58(2):567–574.
31. Rocke DM, et al. A Method for Detection of Differential Gene Expression in the Presence of Inter-Individual Variability in Response. Bioinformatics. In press.
32. Kadhim MA, et al. Transmission of chromosomal instability after plutonium alpha-particle irradiation. Nature 1992; 355(6362):738–740.
33. Kadhim MA, et al. Alpha-particle-induced chromosomal instability in human bone marrow cells. Lancet 1994; 344(8928):987–988.
34. Watson GE, et al. Chromosomal instability in unirradiated cells induced in vivo by a bystander effect of ionizing radiation. Cancer Res 2000;60(20):5608–5611.
35. Whitehouse CA, Tawn EJ. No evidence for chromosomal instability in radiation workers with in vivo exposure to plutonium. Radiat Res 2001;156(5 Pt 1):467–475.
36. Kadhim MA, Marsden SJ, Wright EG. Radiation-induced chromosomal instability in human fibroblasts: Temporal effects and the influence of radiation quality. Int J Radiat Biol 1998;73(2):143–148.
37. Goldberg Z. Clinical implications of radiation-induced genomic instability. Oncogene 2003;22(45):7011–7017.

See also CODES AND REGULATIONS: RADIATION; NONIONIZING RADIATION, BIOLOGICAL EFFECTS OF; RADIATION DOSIMETRY FOR ONCOLOGY; RADIATION PROTECTION INSTRUMENTATION; RADIATION THERAPY, QUALITY ASSURANCE IN.

ION-PAIR CHROMATOGRAPHY. See CHROMATOGRAPHY.

ION–SENSITIVE FIELD-EFFECT TRANSISTORS

PAUL A. HAMMOND
DAVID R.S. CUMMING
University of Glasgow
Glasgow, United Kingdom

INTRODUCTION

The pH of a solution is defined as

$$\mathrm{pH} = -\log[\mathrm{H}^+] \tag{1}$$

where $[\mathrm{H}^+]$ is the concentration of hydrogen ions in the solution. The standard laboratory method of measuring the pH of a solution uses a glass electrode with a thin-walled, bulb-shaped membrane at the bottom. The electrode is filled with a standard solution (usually 0.1 M HCl), into which a silver wire coated with silver chloride is dipped (Fig. 1). Hydrolysis of both the inside and outside of the glass membrane forms thin gel layers, separated by dry glass inside the membrane. Hydrogen ions from the test solution are able to diffuse into the outside gel layer. The glass is doped with mobile ions (e.g., Li^+), which are able to cross the membrane and "relay" the concentration of H^+ ions in the test solution to the inside of the bulb. To make pH measurements, the potential of the internal silver wire is measured with respect to a reference electrode. Since the glass membrane is selective toward $[\mathrm{H}^+]$ ions, the overall cell potential (at room temperature) is given by the Nernst equation:

$$\psi = \mathrm{const} + \frac{kT}{q}\ln[\mathrm{H}^+] \tag{2a}$$

$$= \mathrm{const} + 25.7 \times 10^{-3} \ln 10 \log[\mathrm{H}^+] \tag{2b}$$

$$= \mathrm{const} - 0.0592\,\mathrm{pH} \tag{2c}$$

where k is the Boltzmann constant, T the absolute temperature, and q the electronic charge.

A conventional reference electrode consists of a glass tube containing a filling solution of known composition (e.g., saturated KCl). In an Ag/AgCl electrode, electrical contact is made to the filling solution by a silver wire that has been coated with silver chloride. If saturated KCl is used, the wire develops a potential of 199 mV versus the standard hydrogen electrode (SHE). This potential remains constant as long as the chloride concentration remains constant. The internal filling solution is separated from the test solution by a permeable membrane or "liquid junction", usually made from porous glass or ceramic. Diffusion of ions through the junction provides the conductive path between the reference electrode and the test solution. The KCl is usually used as the filling solution, as the mobility of the K^+ and Cl^- ions are nearly equal, minimising any build-up of junction potential.

The glass-electrode, with its complicated materials and internal reference solution, does not lend itself to miniaturization. However, with the arrival of the metal oxide semiconductor field-effect transistor (MOSFET) in 1960, another method of measuring the interface potential became available. The MOSFET is a three-terminal device in which the voltage on a gate electrode controls the current flowing between source and drain electrodes. The MOSFET has a metal or polysilicon gate electrode, separated from the bulk silicon by a thin, insulating gate oxide layer to provide an extremely high input impedance (Fig. 2a). In an n-channel MOSFET, a positive voltage applied to the gate electrode attracts electrons from the bulk of the p-type silicon to the surface beneath the oxide and creates an inversion region that is rich in mobile electrons. This inversion region forms a channel that allows current to flow between source and drain. The minimum gate voltage

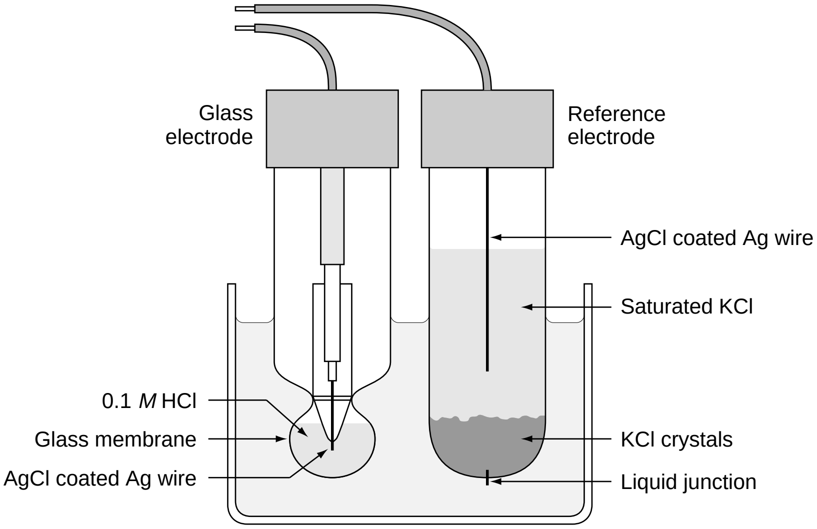


Figure 1. Diagram of the glass electrode together with a silver–silver chloride reference electrode.

that is required to produce "strong" inversion is termed the threshold voltage and is given by

$$V_{\mathrm{T}} = \frac{\Phi_{\mathrm{M}} - \Phi_{\mathrm{S}}}{q} - \frac{Q_{\mathrm{I}}}{C_{\mathrm{I}}} - \frac{Q_{\mathrm{S}}}{C_{\mathrm{I}}} + 2\psi_F \qquad (3)$$

The threshold voltage incorporates the work function difference between the metal gate and the silicon ($\Phi_{\mathrm{M}} - \Phi_{\mathrm{S}}$), any fixed charge in the oxide or in the oxide–silicon interface Q_{I} and the charge in the depletion region of the silicon Q_{S}. The term $2\psi_F$ is twice the Fermi level of the p-type silicon and arises from the definition of strong inversion, that is the inverted material must have a free-electron density that is equivalent to the acceptor density in the p-type material (1). The parameter C_{I} is the capacitance of the oxide (insulator) layer, and all charges and capacitances are expressed per unit area.

In 1970, Bergveld (2) recognized that the MOSFET structure could be adapted to create an ion-sensitive FET (ISFET) by omitting the metal gate. He found that by applying a voltage between the source and drain of this device and placing it in solution, the current flowing would vary with the pH of the solution. This first ISFET used the native gate oxide (SiO_2) as the ion-sensing layer and was sensitive to the concentration of both Na^+ and H^+ ions in the solution (3). The silicon dioxide, used to insulate the metal gate from the silicon substrate, has a similar structure to the permselective membrane used in the glass electrode, and will therefore respond to the concentration of hydrogen ions in a solution. A reference electrode, the electrical contact of which can be regarded as the gate terminal, is used to bias the ISFET (Fig. 2b).

A model for an ISFET, made by excluding the metal gate from a MOSFET, will include all the contributions to its threshold voltage considered in equation 3. There are an additional two terms, one due to the potential of the reference electrode ψ_{REF}, and the other due to the potential at the solution–oxide interface $\Delta\psi + \chi_{\mathrm{SOL}}$. Here, χ_{SOL} is the constant surface dipole potential of the solvent (usually water), and $\Delta\psi$ is the concentration-dependent interface potential. The threshold voltage of the ISFET is then

$$V_{\mathrm{T}} = \psi_{\mathrm{REF}} - \Delta\psi + \chi_{\mathrm{SOL}} - \frac{\Phi_{\mathrm{S}}}{e} - \frac{Q_{\mathrm{I}}}{C_{\mathrm{I}}} - \frac{Q_{\mathrm{S}}}{C_{\mathrm{I}}} + 2\psi_F \qquad (4)$$

since the gate metal of the MOSFET is replaced by the metal in the reference electrode and its work function has been included in ψ_{REF}.

The only term in equation 4 that varies with ionic concentration is $\Delta\psi$, so measuring the concentration is simply a matter of measuring V_{T}. The ISFET is then an ideal transducer with which to measure ionic concentrations since it has an extremely high input impedance, and hence does not require any bias current to flow in the solution. It also uses the same fabrication process as a MOSFET, suggesting that not only can it be made very small, but that it can be integrated on the same substrate as the sensor electronics.

THE SITE-BINDING MODEL

Initially, it was assumed that the silicon dioxide insulator used in the ISFET would obey the Nernst equation, in the same way as the membrane used in the glass electrode did. This meant that, for a pH-ISFET, $\Delta\psi$, and hence V_{T}, would be a linear function of pH with a response of 59 mV·pH^{-1}. The model for the ISFET simply substituted the Nernst equation in place of $\Delta\psi$ in equation 4 (4). While this

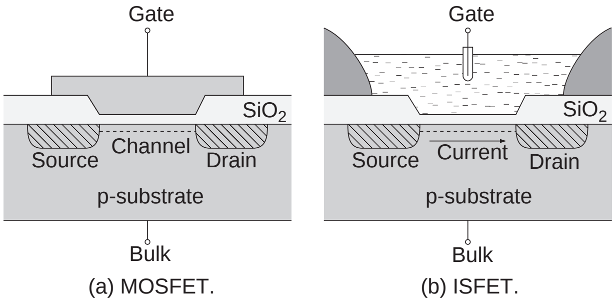


Figure 2. Diagrams showing the similarity in cross-section between a MOSFET and an ISFET.

suggested how to operate the ISFET in a circuit, it did not provide any insight into the chemical processes that occur at the solution–oxide interface. Nor did it provide an explanation for the sub-Nernstian and pH dependent sensitivities that were measured for SiO_2 ISFETs (5).

The so-called site-binding model assumes that the insulator surface has ionizable sites that react directly with the electrolyte to bind or release hydrogen ions. The surface becomes more or less charged depending on the concentration of ions in the solution. The charged surface induces a layer of complementary charge in the solution that forms a double-layer capacitance across which the interface potential is developed. The solution side of the insulator–electrolyte interface is thought to be made up of several "layers" as shown in Fig. 3. The solvent molecules and any ions that are specifically adsorbed onto the surface make up an inner layer. The locus of the electrical centers of the adsorbed ions is called the inner Helmholtz plane (IHP) or Stern layer. The total surface charge density in this plane due to the ions is σ_0. Solvated ions in the solution cannot approach as close to the surface, and their locus of closest approach forms the outer Helmholtz plane (OHP). The interaction of the solvated ions with the surface is purely electrostatic and they are referred to as nonspecifically adsorbed ions. Because of thermal mixing in the solution, these ions are distributed throughout the diffuse layer that extends from the OHP into the bulk of the electrolyte. The excess charge density in the diffuse layer is σ_D so that the total charge in the solution is $\sigma_D + \sigma_0$.

The charge density in the semiconductor region is σ_S, so applying the requirement of charge neutrality:

$$\sigma_D + \sigma_0 + \sigma_S = 0 \tag{5}$$

If the potential in the semiconductor bulk is defined to be zero and the potential of the electrolyte bulk is fixed at ψ_{REF}, then

$$\psi_{REF} + (\psi_D - \psi_{REF}) + (\psi_0 - \psi_D) + (\psi_S - \psi_0) - \psi_S = 0 \tag{6}$$

In addition,

$$\psi_0 - \psi_S = -\frac{\sigma_S}{C_I} \tag{7}$$

$$\psi_0 - \psi_D = -\frac{\sigma_D}{C_H} \tag{8}$$

$$\psi_D - \psi_{REF} = -\frac{2kT}{e}\sinh^{-1}\frac{\sigma_D}{\sqrt{8\varepsilon kTc}} \tag{9}$$

where k is the Boltzmann constant, ϵ is the permittivity, and c is the total ionic concentration of the solution. The last equality (eq. 9) is the Gouy–Chapman model for the diffuse layer (6), and C_H is the capacitance formed between the inner and outer Helmholtz planes. Combining equations 6–9 produces:

$$\psi_{REF} + \underbrace{\frac{-2kT}{e}\sinh^{-1}\left(\frac{\sigma_D}{\sqrt{8\varepsilon kTc}}\right) - \frac{\sigma_D}{C_H}}_{\psi_{EI} \equiv -\Delta\psi} + \underbrace{\frac{\sigma_S}{C_I} - \psi_S}_{\psi_{IS}} = 0 \tag{10}$$

This equation provides a link between the interface potential $\Delta\psi$ and the charge density of the EIS system.

The "site-binding" model was developed by Yates et al. (7) to describe the interactions at a general oxide–electrolyte interface. It was later applied to the electrolyte–SiO_2–Si system by Siu and Cobbold (8) and Bousse et al. (9), whose approach is outlined here. When an SiO_2 surface is in contact with an aqueous solution, it hydrolyzes to form surface silanol (SiOH) groups. These groups are amphoteric, meaning that they can react with either an acid or a base. The acidic and basic character of a neutral SiOH site is described by the following reactions (Fig. 4) and dissociation constants. (The dissociation constant is the equilibrium constant for a reversible dissociation reaction. It expresses the amount by which the equilibrium favors the

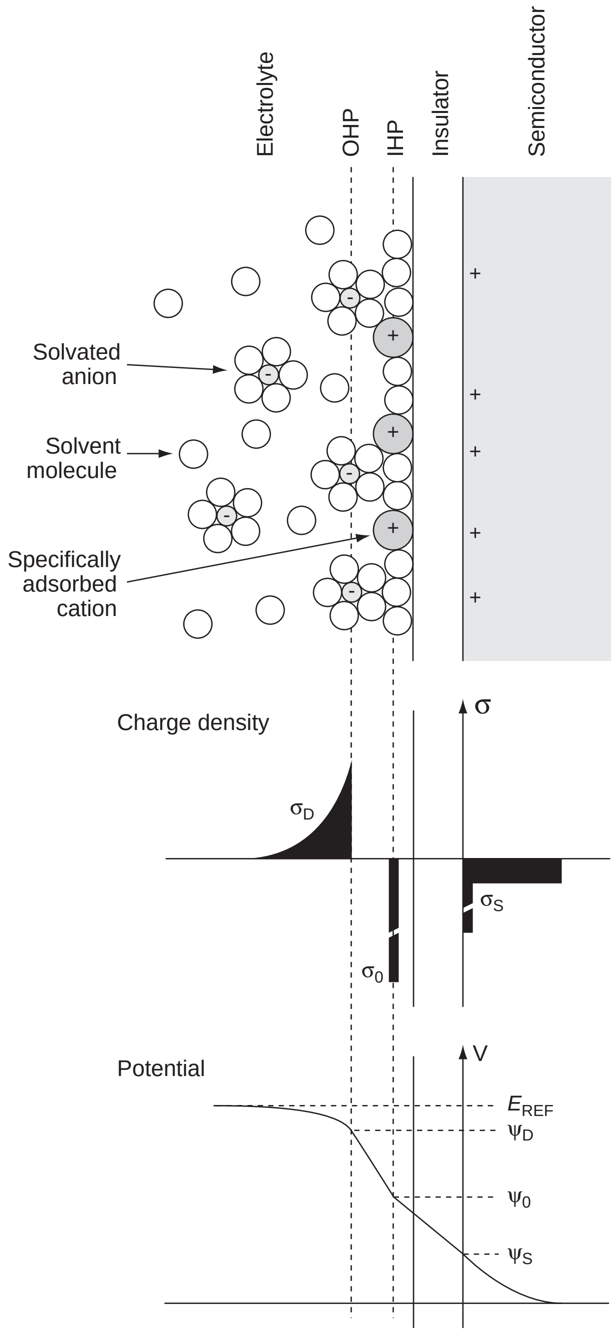


Figure 3. Model of the electrolyte–insulator–semiconductor (EIS) system.

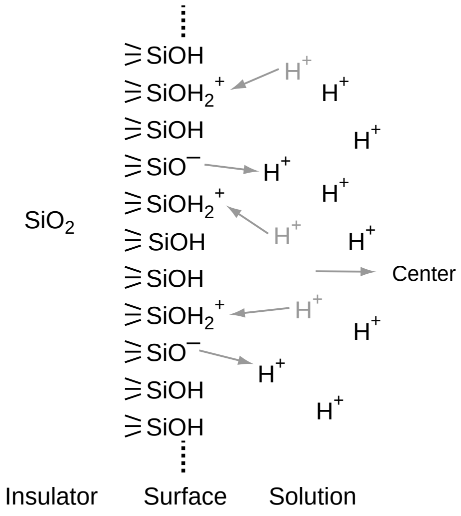


Figure 4. Diagram illustrating the surface sites and hydrogen-ion reactions for SiO_2.

products over the reactants.)

$$SiOH_2^+ \rightleftharpoons SiOH + H^+, \quad K_1 = \frac{[SiOH][H^+]_0}{[SiOH_2^+]} \tag{11}$$

$$SiOH \rightleftharpoons SiO^- + H^+, \quad K_2 = \frac{[SiO^-][H^+]_0}{[SiOH]} \tag{12}$$

where the square brackets indicate concentrations and the subscript 0 is used to indicate a surface quantity. Reactions between surface sites and other ions in the supporting electrolyte (e.g., Na^+, Cl^-) are ignored since they have been shown to have a negligible effect on the interface potential (10).

The density of sites on the surface is

$$N = [SiOH_2^+] + [SiO^-] + [SiOH] \tag{13}$$

and the surface charge per unit area is

$$\sigma_0 = e([SiOH_2^+] - [SiO^-]) \tag{14}$$

Due to thermal mixing, the surface concentration of H^+ ions can be related to the bulk H^+ concentration by Boltzmann statistics:

$$[H^+]_0 = [H^+]\exp\left(\frac{-e\Delta\psi}{kT}\right) \tag{15}$$

as $\Delta\psi$ is the potential difference from the electrolyte bulk to the insulator surface. Multiplying equation 11 by equation 12 and substituting for $[H^+]_0$ using equation 15 gives

$$[H^+] = \sqrt{K_1K_2}\,\exp\left(\frac{e\Delta\psi}{kT}\right)\sqrt{\frac{[SiOH_2^+]}{[SiO^-]}} \tag{16}$$

For the case that $\Delta\psi = 0$ and $\sigma_0 = 0$ $\big($i.e., $[SiOH_2^+] = [SiO^-]\big)$, we can see from equation 16 that $[H^+] = \sqrt{K_1K_2}$. This is the hydrogen ion concentration in the solution required to produce an electrically neutral surface, and is called the point of zero charge (pzc). The pH at this point is denoted pH(pzc) and this can be substituted in 16 to yield

$$2.303(\text{pH(pzc)} - \text{pH}) = \frac{e\Delta\psi}{kT} + \ln\mathcal{F} \tag{17}$$

Table 1. Values for the Parameters of pH Sensitive Insulators Found in the Literature

	K_1	K_2	N_A (sites·m^{-2})	Reference
SiO_2	$10^{1.8}$	$10^{-6.2}$	5×10^{18}	11
Al_2O_3	10^{-6}	10^{-10}	8×10^{18}	12
Ta_2O_3	10^{-2}	10^{-4}	10×10^{18}	12

This equation provides the link between charge, potential, and pH. The function

$$\mathcal{F} = \sqrt{[SiOH_2^+]/[SiO^-]} \tag{18}$$

plays a key role in the response of the surface. It can be written in terms of the "normalized" net charge on the surface $\hat{\sigma}_0 = \sigma_0/eN$, and the parameter $\delta = 2\sqrt{K_2/K_1}$:

$$\mathcal{F} = \frac{\hat{\sigma}_0/\delta + \sqrt{(\hat{\sigma}_0/\delta)^2(1-\delta^2)+1}}{1-\hat{\sigma}_0} \tag{19}$$

Equations 17 and 19 give the solution pH as a function of both $\Delta\psi$ and σ_0, so now it remains to find the relationship between $\Delta\psi$ and σ_0. This is done by using the definition of $\Delta\psi$ in equation 10, the charge neutrality condition in equation 5:

$$\sigma_D + \sigma_0 = \Delta\sigma = -\sigma_S \tag{20}$$

It is usually assumed that $\Delta\sigma = 0$ (9), so that $\sigma_0 = -\sigma_D$. Finally, the interface potential $\Delta\psi$ can be found as a function of the solution pH, by using a parametric method in $\hat{\sigma}_0$. Values of the dissociation constants and surface site density of SiO_2 obtained from the literature are shown in Table 1, and used to generate the SiO_2 pH response curve in Fig. 5. Not only does the SiO_2 surface have a low sensitivity of −46.3 mV/pH (at pH 7), it also has a nonlinear response, especially in the acid pH range.

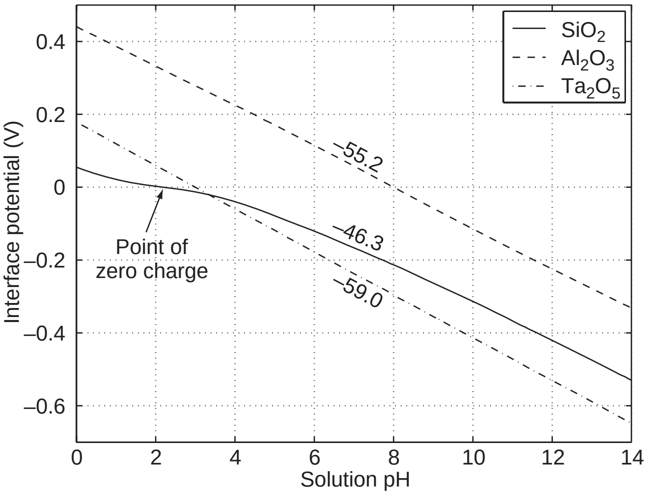


Figure 5. Graph of the theoretical pH response for SiO_2, Al_2O_3, and Ta_2O_5 surfaces.

The insulators Al_2O_3 and Ta_2O_5 also have amphoteric surface groups, so can be modelled using the same parametric method. As shown in Fig. 5, these surfaces produce a linear response with sensitivities much closer to the Nernstian ideal of -59.2 mV·pH^{-1}. As a result, Al_2O_3 and Ta_2O_5 have been widely used as the pH sensitive layer for fabricating ISFETs. The measured sensitivities for these insulators (53–57 mV·pH^{-1} for Al_2O_3, 56–57 mV·pH^{-1} for Ta_2O_5; see Table 2) are close to the theoretical values shown in Fig. 5.

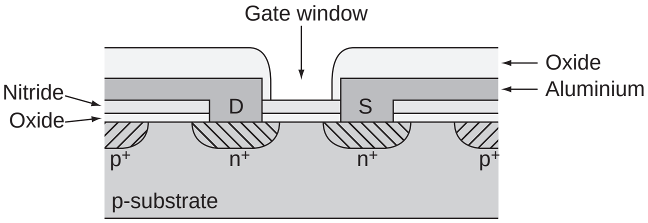


Figure 6. Diagram of the cross-section through an ISFET formed by depositing silicon nitride directly on top of the gate oxide (13).

DEVELOPMENT OF THE ISFET

Much of the subsequent work on ISFETs has concentrated on measuring pH, as this plays a vital role in many biochemical systems. Because silicon dioxide has a low pH sensitivity, the magnitude of which varies with pH (5), Matsuo and Wise (13) experimented with the use of silicon nitride (Si_3N_4) instead. Their ISFET was made by depositing a layer of nitride on top of the thermally grown gate oxide. The ISFET was located at the tip of a needle-shaped probe, which was covered in a thick layer of SiO_2 to insulate it from the solution (Fig. 6). The ISFET was found to have an almost ideal pH response and very low sensitivity to sodium and potassium ion concentrations. Selectivity between ion species is important to differentiate changes in pH from changes in the total ionic concentration of the solution.

In a sense, silicon nitride was an obvious material to investigate as it was, and still is, widely used as a passivation layer to protect devices, such as integrated circuits from the ingress of moisture. Other well-known materials included the oxides of aluminium and tantalum (Al_2O_3 and Ta_2O_5). The pH sensitivity, ion selectivity, response times, and drift rates for these materials have been extensively studied (5). Silicon oxynitride (SiO_xN_y) and other more exotic insulators, such as zirconium and tin oxides (ZrO_2, SnO_2), and even diamond-like carbon (DLC) have all been used to make pH ISFETs. Oxides and nitrides of metals have also been investigated to try and improve parameters, such as response time and maximum operating temperature. These include platinum oxide (PtO_2), titanium nitride (TiN), iridium oxide (Ir_2O_3), and indium tin oxide (ITO). The published pH sensitivity and drift rates (where available) for these materials are summarized in Table 2.

Apart from the choice of pH sensitive material, development of the ISFET has mostly focused on achieving compatibility with the CMOS process. CMOS devices use complementary pairs of n-type and p-type MOSFETs to implement circuits. The CMOS process is the dominant technology for integrated circuits (ICs), so achieving compatibility would allow complex devices containing ISFETs to be fabricated by a standard, industrial process. Figure 7 is a simplified cross-section of a CMOS invertor, showing the polysilicon gate electrodes, the source and drain regions and the metal interconnections.

ISFETs have been made using both NMOS and PMOS transistors. However, in a p-substrate process, the bulk terminal of a PMOS ISFET can be biased above the substrate ground potential, permitting more flexibility in circuit design. The reverse is true for an n-substrate CMOS process. It has also been shown that the noise performance of an ISFET is dominated by $1/f$ or “flicker” noise (23), which is lower in PMOS transistors (24).

Initial attempts to integrate ISFETs involved significant modifications to the standard CMOS process. Wong and White (15) followed the standard sequence of CMOS process steps, until the metallization stage. They then removed the oxide, the polysilicon gate electrode, and the gate oxide above the ISFETs by wet etching. A thinner

Table 2. Values of pH Sensitivity and Drift Rates for ISFETs Made Using a Arange of Materials, Obtained from the Literature

Material	pH Range	Sensitivity, (mV·pH^{-1})	Drift, (mV·h^{-1})	Reference
SiO_2	4–10	25–35 (pH < 7) 37–48 (pH > 7)	Unstable	5
SiO_xN_y	2–8.3	57.4 ± 0.4	0.8 (pH 7)	14
	4–9	18–20	Not mentioned	15
PtO_2	1–10	40.5 ± 4.0	0.5 (pH 6.86)	16
Si_3N_4	1–13	45–56	1.0 (pH 7)	5
ZrO_2	2–10	50	Slow response	17
Al_2O_3	1–13	53–57	0.1–0.2 (pH 7)	5
Ta_2O_5	1–13	56–57	0.1–0.2 (pH 7)	5
DLC	1–12	54–59	3 μV/h (pH 3)	18
SnO_2	2–10	55–58	Not mentioned	19
ITO	2–12	58	Not mentioned	20
TiN	1.68–10.01	59	< 1	21
Ir_2O_3	3–10	59.5	Unclear	22

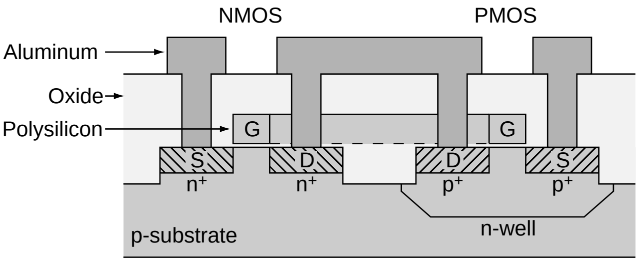


Figure 7. Diagram of the cross-section through a CMOS invertor.

oxide layer was regrown and the sensing layer of Si_3N_4 or Ta_2O_5 was deposited on top of this. The contact windows were then opened and the aluminum deposited to form the interconnects as for a normal CMOS process.

Bousse et al. (25) took a similar approach, but completed the CMOS process before etching away the deposited SiO_2 above the polysilicon gate of the ISFET. They then deposited Si_3N_4 over the whole wafer so that the polysilicon was retained as a floating electrode in the ISFET. This floating electrode did not reduce the ISFET sensitivity to pH, and had the additional benefit of making the ISFET less sensitive to changes in light levels. It does this by shielding the channel from photons that can generate electron–hole pairs, which contribute to the ISFET current. However, since the nitride they used was deposited by low pressure chemical vapor deposition (LPCVD) at 785 °C, the aluminium interconnect had to be replaced with tungsten silicide, which was able to withstand this high temperature step. This meant that a specially modified CMOS process had to be used to fabricate the ISFETs.

Bausells et al. (26) extended the floating electrode idea to a two-metal CMOS process by connecting the polysilicon gate and both metal layers together. In addition, they used the silicon oxynitride passivation layer as the pH sensitive material for the ISFET (Fig. 8). This meant that the ISFET could be fabricated by a commercial foundry using a standard CMOS process, without the need for any process modifications. The fabricated ISFETs had a sensitivity of 47 mV·pH^{-1} and a lifetime of $>$ 2 months. As well as the advantages of using of a well-characterized industrial process, there are additional "system-on-chip" design benefits. These include access to libraries of components such as amplifiers and digital gates that greatly ease the design of the whole system.

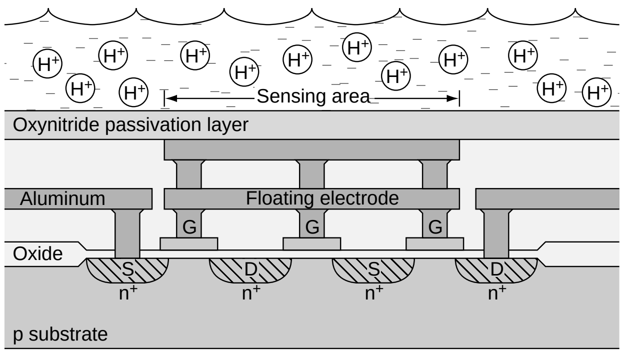


Figure 8. Diagram of the cross-section through an ISFET fabricated by an unmodified commercial CMOS process (26). The floating electrode allows the passivation layer to be used as a pH sensitive insulator.

Unfortunately, the unmodified CMOS ISFETs were found to have large threshold voltages; they were also based on silicon oxynitride, a material that has been found to have a widely varying sensitivity, depending on the deposition conditions (Table 2). The large threshold voltage has been shown to be caused by trapped charge left on the floating electrode during fabrication (27). To avoid these problems, Jakobson et al. (28) removed the passivation layer by using the aluminum of the floating electrode as an etch-stop. They then experimented with low temperature evaporation of pH sensitive layers onto the exposed aluminium. The best performance was obtained by using a layer of platinum (to prevent the aluminum from corroding), followed by a layer of tantalum oxide.

The floating-electrode ISFETs can be considered as special cases of the extended-gate ISFET that was first proposed in 1983. The idea was to separate the electronics from the chemically active region, and by doing so make the device easier to passivate and package than a standard ISFET with an exposed gate insulator. A coaxial polysilicon structure was used to provide a screened connection between the chemically sensitive area and the gate of a MOSFET (29). A more recent CMOS extended-gate ISFET design used a long (unscreened) aluminium track with one end connected to the polysilicon gate and the other exposed by the final pad-etch step (21). This idea has been taken to its limit by using a discrete, off-the-shelf MOSFET and connecting the gate terminal to the pH-sensitive material with a length of wire (20). This method is clearly not applicable to a sensor system-on-chip design, but it does provide a simple method of characterizing the behavior of the material.

The floating-electrode ISFET has also been used to protect against electrostatic discharge (ESD). In the first ESD-protected devices, the polysilicon gate was left intact and connected to a terminal via a MOSFET switch. This provided a reverse-biased diode between the floating electrode and the substrate that would breakdown (reversibly) before the gate insulator was damaged (30). However, current leakage through the "off" MOSFET was such that any response to changing pH decayed to zero in a matter of seconds. To achieve a steady-state response, a large platinum electrode was connected to the ISFET gate to supply current from the solution to replace that being lost though the MOSFET. To avoid the problem of leakage current altogether, ESD-protected ISFETs have been fabricated with a separate platinum ring electrode around the sensitive gate area (31). The platinum electrode is a preferential discharge path to the substrate, protecting the ISFET in the same manner that a lightning conductor protects a building.

ISFETs have also been adapted to create chemically modified FETs (CHEMFETs), which are sensitive to the concentration of ions other than hydrogen. This is achieved by attaching a polymer membrane containing a suitable ionophore to the pH sensing surface of the ISFET. The stability of the ISFET–membrane interface is improved by the addition of an intermediate hydrogel layer. In this way, CHEMFETs sensitive to K^+ (32), Na^+ (33), and other cations have been developed.

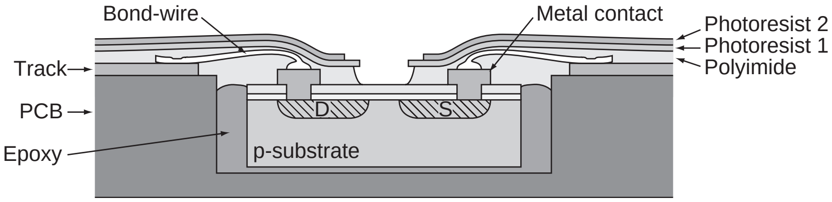


Figure 9. Diagram of the cross-section through an ISFET-based device in a recessed PCB, encapsulated using a layer of polyimide and two layers of photoresist (34).

PACKAGING

One of the main obstacles that has prevented the commercialization of ISFET based devices is the repeatability and reliability of the encapsulation procedure. It is normal for the encapsulant to be applied by hand, covering the chip and bond wires, but leaving a small opening above the sensing area. Epoxy is the most extensively used material although it is important to select a composition that is stable, a good electrical insulator, and does not flow during encapsulation. Many commercially available epoxies have been assessed for their suitability by making measurements of their electrical impedence over time (34–37).

By using ultraviolet (UV) curable polymers, it is possible to increase the automation of the packaging process using a standard mask aligner. A lift-off technique was developed using a sacrificial layer of photosensitive polyimide to protect the ISFET gates. Alumina-filled epoxy was applied by screen printing and partially cured, before the polyimide was etched away leaving a well in the epoxy (38). After 10 days in solution, leakage currents of 200 nA were observed. Better results were achieved by direct photopolymerization of an epoxy-based encapsulant. The ISFETs packaged using this method showed leakage currents of 35 nA after 3 months in solution (38). To avoid polarizing the reference electrode, a leakage current of < 1 nA is desirable (5). This photolithographic patterning of the encapsulant was done at the wafer-level, to all the devices simultaneously. Subsequently, the wafer was diced up and the individual chips were wire-bonded and coated with more encapsulant by hand. At the chip-level, wire-bonded ISFET chips have been covered (again by hand) with a 0.5–1 mm thick photosensitive, epoxy-based film, then exposed and developed (39).

Some degree of automation was introduced by Sibbald et al. (34) who used a dip-coating method to apply the polymers. They first recessed the chip into a PCB and wire-bonded the connections, before coating it with a layer of polyimide. Two layers of photoresist followed, before the underlying polyimide was etched away (Fig. 9). The packaged devices showed < 10 pA leakage current after 10 days in solution. However, the encapsulation did exhibit electrical breakdown for applied bias voltages in excess of 1.5–2 V, which was attributed to the high electric field in the thin layer of resist covering the bond wires. In a separate study, photosensitive polyimide has also been used to create the wells that separate the ion-selective membranes on a multiple ISFET chip (40).

The structure of the ISFET has also been modified to improve the lifetime and ease of manufacture of the packaged device. One solution was to make the ISFET chip long and thin (1.2 × 12 mm) with the sensing area at one end and the bond pads at the other so that the bond-wires did not enter the solution (14). The chip itself was encapsulated with a thick layer of silica. More radical solutions involved bulk micromachining to form back-side contacts to the ISFET so that the bond wires were on the opposite side of the chip to the solution. The front side of the chip is protected by anodic bonding of glass (Fig. 10). A review of back-side contact ISFETs is provided by Cané et al. (41), but the technique is not particularly suited to a CMOS chips, which have many bond-pads arranged around the perimeter.

ISFET CIRCUITS

When an ISFET is placed in solution, a concentration-dependent potential ($\Delta\phi$) is formed at the interface

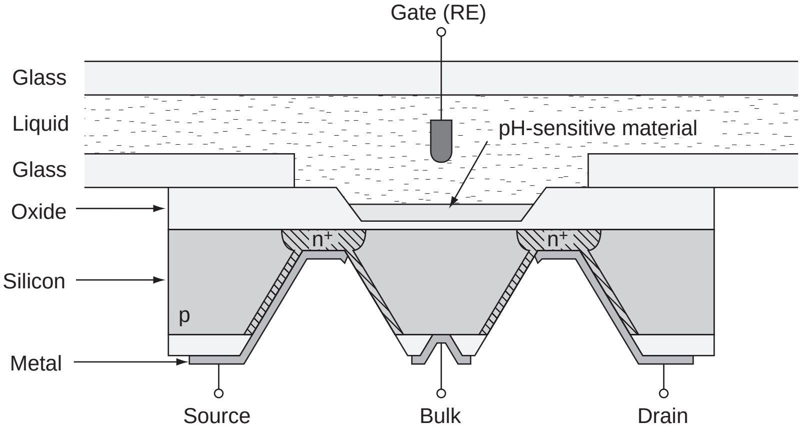
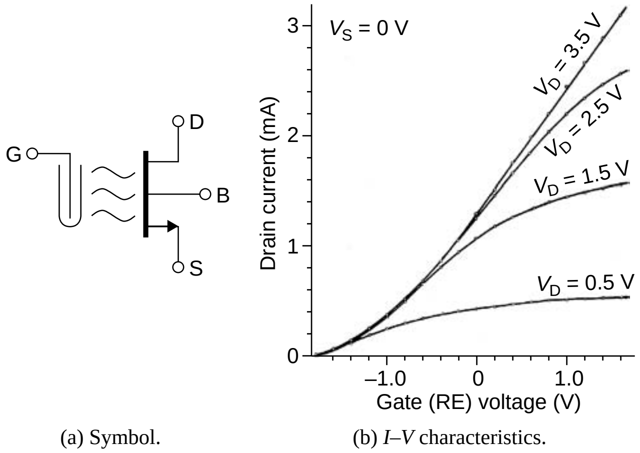


Figure 10. Diagram of the cross-section through a back-side contacted ISFET chip with liquid and electrical contacts on opposite sides (41).



(a) Symbol. (b) *I–V* characteristics.

Figure 11. Diagram of the circuit symbol and a graph of the *I–V* transfer characteristics (4) for an n-type ISFET.

between the gate insulator and the solution. This potential modulates the threshold voltage (V_T) of the ISFET (eq. 4), an internal quantity that cannot be directly measured. Some circuitry is therefore required to convert changes in the threshold voltage into changes of a measurable signal. The solution to this problem lies in the use of feedback control to maintain a constant drain current in the ISFET. For a FET, operating in the linear region[3], the drain current is given by:

$$I_D = k'\frac{W}{L}\left[(V_{GS} - V_T)V_{DS} - \frac{V_{DS}^2}{2}\right] \tag{21}$$

where k' is a process-dependent constant, and W, L are the width, length of the device. (In the linear region, $0 < V_{DS} \leq (V_{GS} - V_T)$ and I_D varies with V_{DS}.) The parameters V_{GS} and V_{DS} are, respectively, the gate-source and drain-source voltages applied to the FET. For an ISFET, a reference electrode in the solution acts as the gate terminal as shown symbolically in Fig. 11a. The I_D vs. V_{GS} curves for an ISFET as measured by Moss et al. (4) are shown in Fig. 11b. It is clear from this graph that biasing an ISFET at a constant drain current is only possible if both V_{DS} and V_{GS} ($= V_G - V_S$) are maintained constant. If a reference electrode is used to control V_G, then from equation 21, as V_T changes with I_D held constant, V_S must change by an equal and opposite amount to compensate. Measuring $\Delta\psi$ (and hence pH) then becomes a straightforward matter of measuring the terminal voltage V_S.

In his original paper on the operation of the ISFET, Bergveld (3) stated that one of the important advantages of ISFETs, compared with conventional pH electrodes, is that there is no need for a reference electrode. Instead, he used a feedback circuit to control the bulk terminal of the ISFET and maintain a constant drain current. However, this makes the assumption that the solution is perfectly isolated from the ISFET source and drain terminals, as well as the circuit. Any current (even leakage current through the packaging), that flows into the solution will affect its potential. To a bulk-feedback circuit, this will be indistinguishable from a change in solution concentration. It is therefore safer to assume that the solution is grounded, or at least at some well-defined potential with respect to the circuit, and use a reference electrode to ensure that this is the case. For this reason, all of the subsequently published circuits relating to ISFETs include a reference electrode.

The probe fabricated by Matsuo and Wise (13) contained an ISFET and a MOSFET of identical dimensions. The ISFET was configured as a source follower with the MOSFET acting as a constant current source. A saturated calomel electrode (SCE, shown in Fig. 1) was used as a reference, and the output voltage measured at the source terminal of the ISFET. Bergveld (42) used a grounded reference electrode to avoid the problem of a short circuit if the solution is already at ground potential (e.g., in an earthed metal container). He used an instrumentation amplifier arrangement to maintain a constant current at a constant drain-source voltage (Fig. 12). Amplifiers A_1 and A_2 set V_{DS} as determined by the fixed current flowing in R_1. Current feedback from amplifier A_4 adjusts the drain voltage via R_2 to keep the current constant as the threshold voltage changes. One disadvantage of this circuit is that the output (measured across R_9) is not referenced to a fixed voltage such as ground.

There have been many other circuit topologies proposed to keep the drain current and/or the drain-source voltage constant. A straightforward circuit that achieves both of

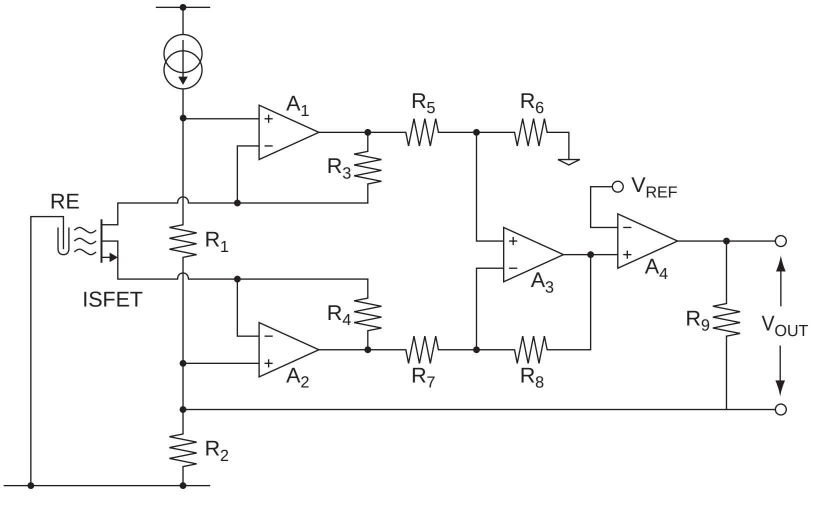


Figure 12. Circuit diagram of an ISFET source and drain follower circuit used to maintain a constant drain current at a constant drain-source voltage (42).

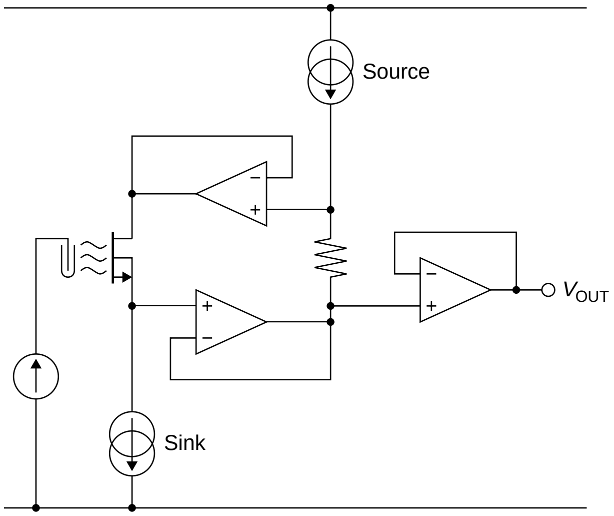


Figure 13. Circuit diagram of constant drain current, constant drain–source voltage ISFET bias circuit (43).

these objectives was presented by Ravezzi and Conci (43). It used a pair of unity-gain, noninverting operational amplifiers (op-amps) to ensure that the voltage dropped across a resistor is also dropped across the ISFET (Fig. 13). Constant current was maintained in the resistor by a current source, and in the ISFET by a current sink. This arrangement allows the source and drain potentials to adjust as the threshold voltage changes, allowing the reference electrode to remain at a fixed potential.

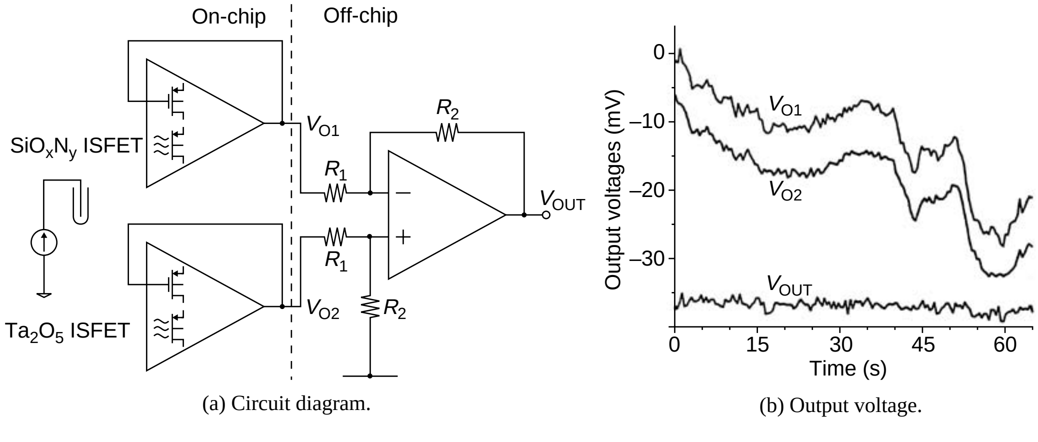
The first integrated ISFET circuit was the so-called operational transducer published by Sibbald (44) in 1985. It used an ISFET and a MOSFET with identical geometries as the active devices in the "long-tailed pair" input stage of an amplifier. Feedback was used to control the gate voltage of the MOSFET to ensure that the same current flowed in both devices. The MOSFET gate voltage tracked that of the ISFET and so measured the changes in threshold voltage. The key advantage of this circuit was that changes in temperature affected both devices equally and were canceled out.

Wong and White (15) recognized that there was little to be gained from an integrated, miniaturized sensor if it relied on a large, external reference electrode. Instead, they used an on-chip gold contact as a quasi-reference electrode (qRE) and a differential circuit to make measurements between two ISFETs with different pH sensitivities. The potential difference between the solution and the qRE will depend on the solution composition. However, like temperature, this is a common-mode signal, which will affect both ISFETs equally. Hence, it can be rejected by means of a differential circuit. Tantalum oxide and silicon oxynitride were used as sensing layers for the two ISFETs, which formed the input stages of op-amps integrated onto a CMOS chip (Fig. 14a). The outputs from the two op-amps were fed into an off-chip differential amplifier. The overall circuit gave a response of 40–43 mV·pH^{-1}. The benefit of the differential approach can be seen in Fig. 14b, which shows the single-ended (V_{O1} and V_{O2}) and differential (V_{OUT}) responses as electrical noise was deliberately applied to the solution. Two copies of the bias circuit in Fig. 13 have also been used to create a differential system with Si_3N_4 and SiO_2 ISFETs (45).

The concept of integration has been extended by Hammond et al. (46) who created a complete digital pH meter on a single CMOS chip. This design makes use of the libraries of components provided by the CMOS foundry to integrate not only the ISFET, but also analog bias circuits, digital signal processing, and storage onto the same chip (Fig. 15a). The chip was mounted in a recessed PCB and covered with a thick layer of photoresist so that only the ISFET area was exposed. The digital response of the device to the changing pH of the solution in which it is immersed is shown in Fig. 15b.

MINIATURE REFERENCE ELECTRODES

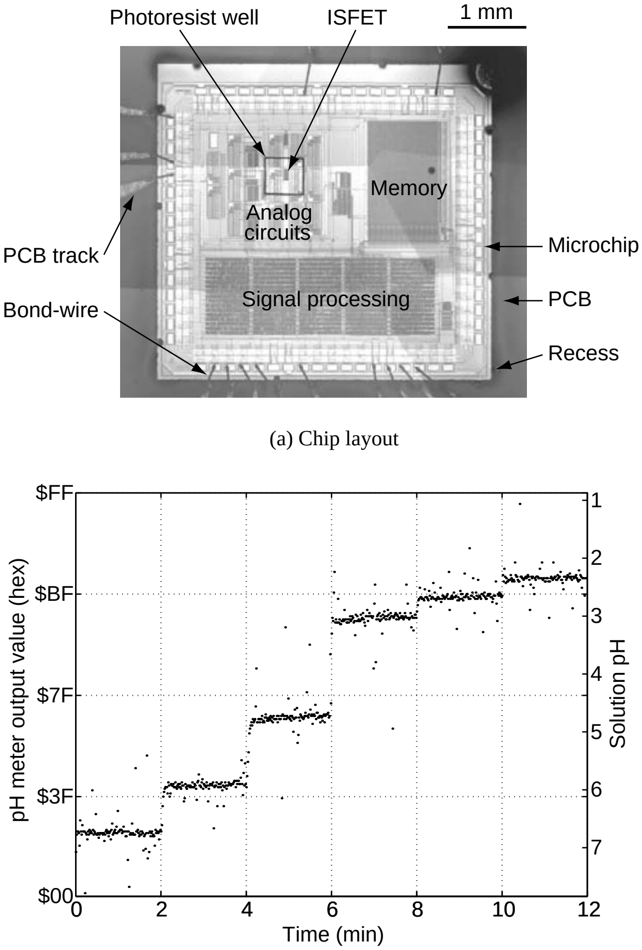
The first attempt to incorporate a reference electrode on an ISFET chip used a thin-film Ag/AgCl electrode (47). Electrodes like this, with no internal reference solution, are sensitive to changes in concentration of their primary ion (in this case Cl$^-$), and are referred to as quasi-reference electrodes. To solve this problem, Smith and Scott (48) also



(a) Circuit diagram.

(b) Output voltage.

Figure 14. Circuit diagram and graph of output voltage for a differential integrated amplifier based on Ta_2O_5 and SiO_xN_y ISFETs (15).



(a) Chip layout

(b) pH response

Figure 15. Chip layout and pH response of a single-chip digital pH meter.

integrated the reference solution and porous membrane into the chip. Wells were etched into the back-side of a silicon wafer, to leave a membrane 10–70 μm thick. The membranes were anodized to porous silicon in a bath of concentrated hydrofluoric acid. The wells were filled with saturated KCl and sealed with glass coverslips that had been coated with thin films of Ag/AgCl. The reference electrode exhibited a low drift rate of 80 $\mu V \cdot h^{-1}$ (worst-case) and a lifetime of > 2 weeks. However, a method of mass producing an integrated, liquid-filled electrode has yet to be developed.

Recent developments of miniature reference electrodes have focused on the use of polymer-supported electrolyte gels, to replace the liquid filling solution. Suzuki et al. (49) developed an electrode that uses finely ground KCl powder supported in a matrix of poly(vinylpyrrolidone) (PVP). An exploded diagram of the electrode is shown in Fig. 16. First, a layer of silver was evaporated onto a glass substrate, inside a U-shaped gold backbone. A layer of polyimide was applied to protect the silver and to define the electrode structure. The AgCl was grown through a slit in the polyimide, and a liquid junction was formed by casting a hydrophilic polymer into the square recess. The electrolyte layer, containing the KCl powder, was then screen-printed over the AgCl and the liquid junction. Finally, a passivating layer of silicone rubber was applied. The electrode can be stored dry, and activated when required by the injection of a saturated solution of KCl and AgCl through the silicone. This miniature reference electrode showed a stability of ±1.0 mV over a period of 100 h. No difference was observed between experimental data obtained with the miniature reference electrode and with a large, commercial reference electrode.

REFERENCE FETs

The sensitivity of the differential circuits already discussed can be increased if one of the devices has no response to

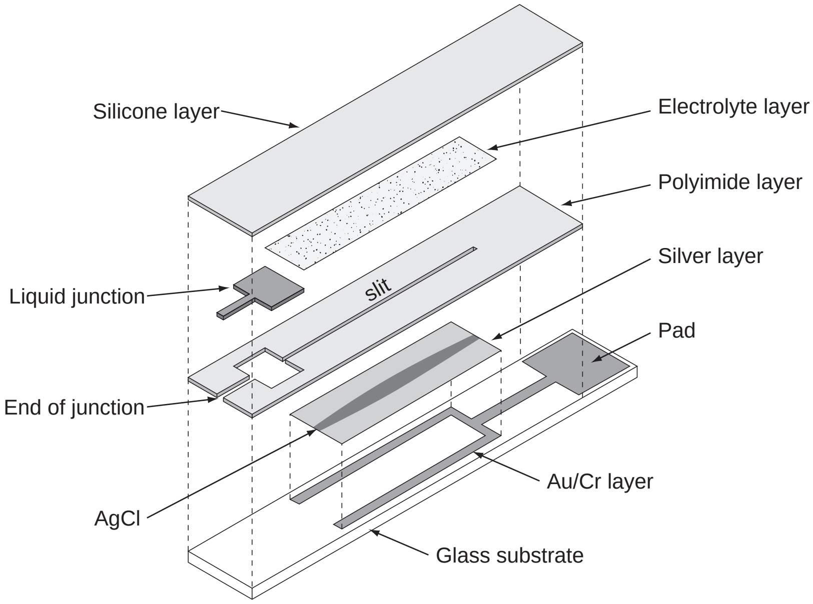


Figure 16. Exploded diagram showing the construction of a miniature Ag/AgCl reference electrode based on a polymer-supported electrolyte gel (49).

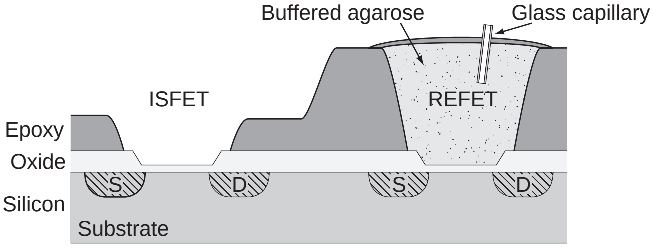


Figure 17. Diagram of the cross-section through a REFET created using encapsulated pH buffer solution (50).

changes in pH. Such a device is called a reference FET (REFET). The first REFET was presented by Comte and Janata (50) in 1978. It consisted of an ISFET surrounded by a well of epoxy that was filled with a buffered agarose gel. A glass capillary was inserted into the gel and sealed in place with more epoxy (Fig. 17). This acted as a liquid junction between the internal reference gel and the external solution. The response of the REFET was only 3 mV·pH^{-1} and it provided temperature compensation of ±0.01 pH·C^{-1} when used in a differential arrangement. However, the techniques required to prepare this REFET are not well suited to mass production.

The pH sensitivity of an ISFET is due to the presence of chemical sites on the surface of the insulator (Fig. 4) that can exchange hydrogen ions with the solution. Attempts to reduce the density of these sites, and hence the sensitivity, by chemical modification of the surface proved unsuccessful (51). Instead a thin membrane of parylene was used to cover an ISFET and convert it into a REFET (52). Parylene has an extremely low density of surface sites and the REFET was found to have a very low pH sensitivity of only 0.5 mV·pH^{-1}. However, parylene forms an insulating (or ion-blocked) layer that affects the electrical properties of the underlying ISFET. Even a very thin membrane reduces the gate capacitance, and hence transconductance, dramatically. This is a problem for differential circuits, which rely on ISFET and REFET having well-matched electrical properties. If a membrane could be found whose ion-conducting properties were sufficient to pass the electrical voltage of the solution to the sensing layer of the underlying ISFET, the transconductance would not be changed. Clearly, such "ion-unblocking" membranes must be insensitive to variations in pH.

Bergveld et al. (53) investigated several candidate polymers for REFET membranes and found that polyacrylate gave the best performance. An intermediate layer of buffered poly(hydroxyethyl methacrylate) (p-HEMA) hydrogel was necessary to fix the interface potential and to improve the adhesion of the membrane (54). The REFET showed < 2 mV·pH^{-1} sensitivity, good mechanical properties, and its transconductance matched that of the ISFET. Despite these useful properties, the acrylate membrane was selectively permeable for cations (e.g., potassium ions). This problem was solved by optimizing the amount of didodecyldimethylammonium bromide (DDMAB) added to the membrane (Fig. 18a). This large immobile cation repels mobile cations (e.g., potassium), from the membrane. Figure 18b shows the performance of the ISFET–REFET differential system, as compared to the individual devices, measured using a platinum qRE.

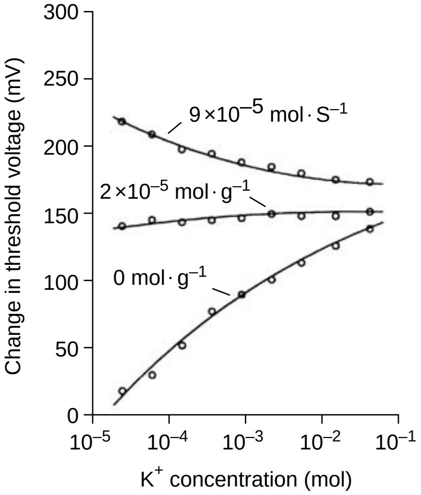


Figure 18. Graphs of the response of an acrylate REFET, and an ISFET, to changes in potassium and hydrogen ion concentrations (53).

An alternative approach was developed by van den Vlekkert et al. (55) who used a thick layer of p-HE MA hydrogel to produce an ISFET with a retarded pH response. By controlling the diffusion coefficient of hydrogen ions, a pseudo-REFET with a nonresponse time of ∼ 10 min was created. Such a REFET is only really useful in a flow-through system where the analyte is pumped through in short bursts followed by rinsing and calibration solutions (34). The p-HEMA hydrogel was also used by Chudy et al. (56) as the base layer for chemically sensitive FETs (CHEMFETs) and REFETs. The CHEMFET uses an additional ion-selective membrane containing an ionophore, selected to target the ion of interest. Exclusion of the ionophore from the membrane enabled the creation of a REFET that showed almost no response to pH, potassium, or calcium ions.

APPLICATIONS

According to Bergveld (57), ∼ 20 companies have commercialised ISFETs based on 150 patents. Product offerings and specifications vary, but in general terms ISFETs have found applications where small size, fast response, robustness, wide operating temperature range, and operation in nonaqueous environments are desirable. The ISFET can also be stored dry, making it more convenient to use than traditional glass electrodes, since the lifetime of a stored ISFET is almost indefinite. ISFETs are used for high resolution applications and products are available with specified resolutions < 0.01 pH units. However, many of the devices on the market have a resolution of ∼ 0.1 pH units.

The ISFET has made a significant impact on a number of industries where it is a requirement to monitor and control the addition of reagents. It made an early appearance in the food industry where its robust construction allowed measurement to be made in foodstuffs. In addition, its ability to operate in nonaqueous environments is an advantage when working with meat and dairy products. The ISFET has also found applications in the pharmaceutical industry, where a fast re-sponse and high operating temperatures are required. It has also been used in electroplating where it is able to withstand the corrosive plating solutions. In addition to manufacturing industries, the ISFET is also used in waste water management and in the treatment of effluent.

Early researchers considered medical applications ranging from monitoring the pH in the mouth (particularly in the context of tooth decay), to direct measurement of ionic balance in the blood. ISFETs been built into dental prosthetics to enable the direct measurement of pH in the presence of plaque (58). Recent data on tooth decay was obtained using a pH imaging microscope (59) although, unlike the ISFET device, this does not allow for *in situ* observations to be made. ISFETs are also used indirectly for dental applications, for example, in the evaluation of potential prosthetic materials (60). The ISFET has been modified in a manner similar to that for the REFET for use in blood analysis beyond measuring ionic concentrations. In this embodiment, the ISFET can be made into a sensitive enzyme sensor. In their paper, Sant et al. (61) demonstrated a sensor for creatinine based on an ISFET with a sensitivity of 30 mV·pCreatinine^{-1}. The application is in renal failure and haemodialysis in particular.

Recently, there has been a growth of interest in the use of ISFETs in a device known as a diagnostic pill. The concept of a diagnostic pill was developed in the 1950s and 1960s (62). Such devices are small enough to be swallowed by a patient, and once inside the gastrointestinal tract, can measure and wirelessly transmit data to an external receiver over a period of many hours or even days. The earliest devices were used to measure pressure changes in the gut, but glass-electrode-based pills were also made to measure gut pH. The diagnostic pill has made something of a comeback in recent years, particularly with the invention of the video pill (63), a device that is capable of wirelessly transmitting images of the gut with considerably less patient discomfort than would be caused by an endoscopic procedure. Various new examples of pH measuring pills have also been developed. The Bravo capsule, which does not actually use an ISFET, is designed for use in the esophagus (64). During operation it is pinned to the lining of the esophagus so that it can monitor conditions such as gastro-esophageal reflux disease (GERD) over a period of 2–3 days. The IDEAS capsule (65) that does use an ISFET, has been built for use in the lower gastrointestinal tract. The device is designed pass naturally through the gut with as little intervention as possible and is required to be able to operate in difficult conditions where biofouling of the sensor is a potential problem.

CONCLUSIONS

The ISFET first appeared in 1970 as an offshoot of the rapidly growing capability of the microelectronics industry. In its simplest form, the ISFET is a modification of the traditional MOSFET in which the gate electrode is removed and the gate oxide exposed to solution. The ISFET uses changes in the surface chemistry of the gate oxide to modify the threshold voltage of the transistor and produce an electronic signal. The sensitivity and performance of the ISFET is highly dependent on the material used to form the gate oxide layer. The behavior of such materials is best modeled using a site-binding approach to calculate the variation in surface charge with solution pH. A wide variety of materials have been experimented with, those found to give the best performance are the oxides of aluminium and tantalum. However, in a standard CMOS manufacturing process, the passivation layer is made of silicon nitride, which is also a good pH sensitive material. It is therefore possible to make good ISFETs using a standard foundry process with little or no additional effort. As a result it has become possible to implement integrated ISFET circuits. A number of circuit topologies for detecting the change in ISFET threshold voltage have been designed, those using voltage followers to maintain the ISFET bias conditions produce the best results. Further development of ISFET integrated circuits has enabled complete instruments to be fabricated on a single IC. To avoid the use of a bulky reference electrode, differential circuits using matched pairs of ISFET and REFET have been designed. However, difficulties in creating a good REFET have increased interest in developing miniature reference electrodes that are compatible with IC processing. The ISFETs have already found widespread application in manufacturing industries, environmental monitoring and medicine. It is expected that with improved miniaturization, integration, and packaging technologies, new applications will emerge.

BIBLIOGRAPHY

1. Streetman BG, Banerjee S. Solid State Electronic Devices. 5th ed. New Jersey: Prentice Hall; 2000.
2. Bergveld P. Development of an ion-sensitive solid-state device for neurophysiological measurements. IEEE Trans Biomed Eng 1970;BM17:70.
3. Bergveld P. Development, operation, and application of ion-sensitive field effect transistor as a tool for electrophysiology. IEEE Trans Biomed Eng 1972;BM19:342.
4. Moss SD, Johnson CC, Janata J. Hydrogen, calcium, and potassium ion-sensitive FET transducers—preliminary-report. IEEE Trans Biomed Eng 1978;25:49–54.
5. Matsuo T, Esashi M. Methods of ISFET fabrication. Sens Actuator 1981;1:77–96.
6. Bard AJ, Faulkner LR. Electrochemical Methods—Fundamentals and Applications. Hoboken (NJ), John Wiley; 2001.
7. Yates DE, Levine S, Healy TW. Site-binding model of electrical double-layer at oxide-water interface. J Chem Soc Faraday Trans 1974;70:1807–1818.
8. Siu WM, Cobbold RSC. Basic properties of the electrolyte-SiO_2-Si system-physical and theoretical aspects. IEEE Trans Electron Devices 1979;26:1805–1815.
9. Bousse L, de Rooij NF, Bergveld P. Operation of chemically sensitive field-effect sensors as a function of the insulator-electrolyte interface. IEEE Trans Electron Devices 1983;30: 1263–1270.
10. Bousse L, de Rooij NF, Bergveld P. The influence of counterion adsorption on the ψ_0/pH characteristics of insulator surfaces. Surf Sci 1983;135:479–496.

11. Harame DL, Bousse LJ, Shott JD, Meindl JD. Ion-sensing devices with silicon-nitride and borosilicate glass insulators. IEEE Trans Electron Devices 1987;34:1700–1707.
12. van Hal REG, Eijkel JCT, Bergveld P. A general model to describe the electrostatic potential at electrolyte oxide interfaces. Adv Colloid Interface Sci 1996;69:31–62.
13. Matsuo T, Wise KD. Integrated field-effect electrode for biopotential recording. IEEE Trans Biomed Eng 1974;BM21: 485–487.
14. Rocher V, et al. An oxynitride ISFET modified for working in a differential-mode for pH detection. J Electrochem Soc 1994; 141:535–539.
15. Wong HS, White MH. A CMOS-integrated ISFET-operational amplifier chemical sensor employing differential sensing. IEEE Trans Electron Devices 1989;36:479–487.
16. Tsukada K, Miyahara Y, Miyagi H. Platinum-platinum oxide gate pH ISFET. Jpn J Appl Phys Part 1-Regul Pap Short Notes Rev Pap 1989;28:2450–2453.
17. Akiyama T, et al. Ion-sensitive field-effect transistors with inorganic gate oxide for pH sensing. IEEE Trans Electron Devices 1982;29:1936–1941.
18. Voigt H, et al. Diamond-like carbon-gate pH-ISFET. Sens Actuator B-Chem 1997;44:441–445.
19. Liao HK, et al. Study of amorphous tin oxide thin films for ISFET applications. Sens Actuator B-Chem 1998;50:104–109.
20. Yin LT, et al. Study of indium tin oxide thin film for separative extended gate ISFET. Mater Chem Phys 2001;70:12–16.
21. Chin YL, et al. Titanium nitride membrane application to extended gate field effect transistor pH sensor using VLSI technology. Jpn J Appl Phys Part 1-Regul Pap Short Notes Rev Pap 2001;40:6311–6315.
22. Katsube T, Lauks I, Zemel JN. pH-sensitive sputtered iridium oxide-films. Sens Actuator 1982;2:399–410.
23. Jakobson CG, Nemirovsky Y. 1/f noise in ion sensitive field effect transistors from subthreshold to saturation. IEEE Trans Electron Devices 1999;46:259–261.
24. Jolly RD, McCharles RH. A low-noise amplifier for switched capacitor filters. IEEE J Solid-State Circuit 1982;17:1192–1194.
25. Bousse L, Shott J, Meindl JD. A process for the combined fabrication of ion sensors and CMOS circuits. IEEE Electron Device Lett 1988;9:44–46.
26. Bausells J, Carrabina J, Errachid A, Merlos A. Ion-sensitive field-effect transistors fabricated in a commercial CMOS technology. Sens Actuator B-Chem 1999;57:56–62.
27. Hammond PA, Ali D, Cumming DRS. Design of a single-chip pH sensor using a conventional 0.6 μm CMOS process. IEEE Sens J 2004;4:706–712.
28. Jakobson CG, Dinnar U, Feinsod M, Nemirovsky Y. Ion-sensitive field-effect transistors in standard CMOS by post processing. IEEE Sens J 2002;2(4):279–287.
29. van der Spiegel J, Lauks I, Chan P, Babic D. The extended gate chemically sensitive field-effect transistor as multi-species microprobe. Sens Actuator 1983;4:291–298.
30. Smith R, Huber RJ, Janata J. Electrostatically protected ion sensitive field-effect transistors. Sens Actuator 1984;5: 127–136.
31. Baldi A, Bratov A, Mas R, Domínguez C. Electrostatic discharge sensitivity tests for ISFET sensors. Sens Actuator B-Chem 2001;80:255–260.
32. Reinhoudt DN, et al. Development of durable K^+-selective chemically-modified field-effect transistors with functionalized polysiloxane membranes. Anal Chem 1994;66:3618–3623.
33. Brunink JAJ, et al. Chemically modified field-effect transistors - a sodium-ion selective sensor based on calix[4]arene receptor molecules. Anal Chim Acta 1991;254:75–80.
34. Sibbald A, Whalley PD, Covington AK. A miniature flow-through cell with a 4-function CHEMFET integrated-circuit for simultaneous measurements of potassium, hy-drogen, calcium and sodium-ions. Anal Chim Acta 1984;159:47–62.
35. Chovelon JM, Jaffrezic-Renault N, Fombon JJ, Pedone D. Monitoring of ISFET encapsulation aging by impedance measurements. Sens Actuator B-Chem 1991;3:43–50.
36. Grisel A, Francis C, Verney E, Mondin G. Packaging technologies for integrated electrochemical sensors. Sens Actuator 1989;17:285–295.
37. Gràcia I, Cané C, Lora-Tamayo E. Electrical characterization of the aging of sealing materials for ISFET chemical sensors. Sens Actuator B-Chem 1995;24:206–210.
38. Muñoz J, et al. Planar compatible polymer technology for packaging of chemical microsensors. J Electrochem Soc 1996;143:2020–2025.
39. Bratov A, Muñoz J, Dominguez C, Bartrolí J. Photocurable polymers applied as encapsulating materials for ISFET production. Sens Actuator B-Chem 1995;25:823–825.
40. Tsukada K, Sebata M, Miyahara Y, Miyagi H. Long-life multiple-ISFETs with poly-meric gates. Sens Actuator 1989;18: 329–336.
41. Cané C, Gràcia I, Merlos A. Microtechnologies for pH ISFET chemical sensors. Micro-electron J 1997;28:389–405.
42. Bergveld P. The operation of an ISFET as an electronic device. Sens Actuator 1981;1:17–29.
43. Ravezzi L, Conci P. ISFET sensor coupled with CMOS read-out circuit microsystem. Electron Lett 1998;34:2234–2235.
44. Sibbald A. A chemical-sensitive integrated-circuit - the operational transducer. Sens Actuator 1985;7:23–38.
45. Palán B, et al. New ISFET sensor interface circuit for biomedical applications. Sens Actuator B-Chem 1999;57:63–68.
46. Hammond PA, Ali D, Cumming DRS. A system-on-chip digital pH meter for use in a wireless diagnostic capsule. IEEE Trans Biomed Eng 2005;52:687–694.
47. Harame DL, Shott JD, Bousse L, Meindl JD. Implantable ion-sensitive transistors. IEEE Trans Biomed Eng 1984;31:572–572.
48. Smith RL, Scott DC. An integrated sensor for electrochemical measurements. IEEE Trans Biomed Eng 1986;33:83–90.
49. Suzuki H, Shiroishi H, Sasaki S, Karube I. Microfabricated liquid junction Ag/AgCl reference electrode and its application to a one-chip potentiometric sensor. Anal Chem 1999;71:5069–5075.
50. Comte PA, Janata J. Field-effect transistor as a solid-state reference electrode. Anal Chim Acta 1978;101:247–252.
51. van den Berg A, Bergveld P, Reinhoudt DN, Sudhölter EJR. Sensitivity control of ISFETs by chemical surface modification. Sens Actuator 1985;8:129–148.
52. Matsuo T, Nakajima H. Characteristics of reference electrodes using a polymer gate ISFET. Sens Actuator 1984;5:293–305.
53. Bergveld P, et al. How electrical and chemical requirements for REFETs may coincide. Sens Actuator 1989;18:309–327.
54. Sudhölter EJR, et al. Modification of ISFETs by covalent anchoring of poly(hydroxyethyl methacrylate) hydrogel-introduction of a thermodynamically defined semiconductor-sensing membrane interface. Anal Chim Acta 1990;230:59–65.
55. van den Vlekkert HH, de Rooij NF, van den Berg A, Grisel A. Multi-ion sensing system based on glass-encapsulated ph-ISFETs and a pseudo-REFET. Sens Actuator B-Chem 1990;1:395–400.
56. Chudy M, Wróblewski W, Brzózka Z. Towards REFET. Sens Actuator B-Chem 1999;57:47–50.
57. Bergveld P. Thirty years of ISFETOLOGY-what happened in the past 30 years and what may happen in the next 30 years. Sens Actuator B-Chem 2003;88:1–20.
58. Visch LL, Bergveld P, Lamprecht W, Sgravenmade EJ. pH measurements with an ion sensitive field-effect transistor in

the mouth of patients with xerostomia. IEEE Trans Biomed Eng 1991;38:353–356.

59. Hiraishi N, et al. Evaluation of active and arrested carious dentin using a pH-imaging microscope and an X-ray analytical microscope. Oper Dent 2003;28:598–604.
60. De Aza PN, Luklinska ZB, Anseau M. Bioactivity of diopside ceramic in human parotid saliva. J Biomed Mater Res Part B 2005;73B:54–60.
61. Sant W, et al. Development of a creatinine-sensitive sensor for medical analysis. Sens Actuator B-Chem 2004;103:260–264.
62. Mackay RS. Radio telemetering from within the body. Science 1961;134:1196–1202.
63. Iddan G, Meron G, Glukhovsky A, Swain P. Wireless capsule endoscopy. Nature (London) 2000;405:417–417.
64. Pandolfino JE, et al. Ambulatory esophageal pH monitoring using a wireless system. Am J Gastroenterol 2003;98:740–749.
65. Johannessen EA, et al. Implementation of multichannel sensors for remote biomedical measurements in a microsystems format. IEEE Trans Biomed Eng 2004;51:525–535.

See also IMMUNOLOGICALLY SENSITIVE FIELD-EFFECT TRANSISTORS; INTEGRATED CIRCUIT TEMPERATURE SENSOR.

ISFET. See ION-SENSITIVE FIELD-EFFECT TRANSISTORS.

JOINTS, BIOMECHANICS OF

George Papaioannou
University of Wisconsin
Milwaukee, Wisconsin

Yener N. Yeni
Henry Ford Hospital
Detroit, Michigan

INTRODUCTION

The human skeleton is a system of bones joined together to form segments or links. These links are movable and provide for the attachment of muscles, ligaments, tendons, and so on. to produce movement. The junction of two or more bones is called an articulation. There are a great variety of joints even within the human body and a multitude of types among living organisms that use exo- and endoskeletons to propel. Articulation can be classified according to function, position, structure and degrees of freedom for movement they allow, and so on. Joint biomechanics is a division of biomechanics that studies the effect of forces on the joints of living organisms.

Articular Anatomy, Joint Types, and Their Function

Anatomic and structural classification of joints typically results in three major categories, according to the predominant tissue or design supporting the articular elements together, that is, joints are called fibrous, cartilaginous, or synovial.

Synovial joints are cavitated. In general, two rigid skeletal segments are brought together by a capsule of connective tissue and several other specialized tissues, that form a cavity. The joints of the lower and upper limbs are mainly synovial since these are the most mobile joints. Mobility varies considerably and a number of subcategories are defined based on the specific shape or architecture and topology of the surfaces involved (e.g., planar, saddle, ball and socket) and on the types of movement permitted (e.g., flexion and extension, medial and lateral rotation) (Table 1). The basic structural characteristics that define a synovial joint can be summarized in four features: a fibrous capsule that forms the joint cavity, a specialized articular cartilage covering the articular surfaces, a synovial membrane lining the inner surface of the capsule that also secretes a special lubricating fluid, the synovial fluid. Additional supportive structures in synovial joints include disks, menisci, labra, fat pads, tendons, and ligaments.

Cartilaginous joints are also solid and are more commonly known as synchondroses and symphyses, a classification based on the structural type of cartilage that intervenes between the articulating parts (Table 2). This cartilage is hyaline and fibrocartilage for synchondroses and *symphyses*, respectively. *Synchondroses* allow very little movement as in the case of the rib cage that contributes to the ability of this area to expand with respiration. Most *symphyses* are permanent; those of sacrum and coccyx can, however, degenerate with subsequent fusion between adjacent vertebral bodies as part of the normal development of these bones.

Fibrous joints are solid. The binding mechanism that dominates the connectivity of the articulating elements is principally fibrous connecting tissue, although other tissue types also may be present. Length, specific arrangement, and fiber density vary considerably according to the location of the joint and its functional requirements. Fibrous joints are classified in three groups: sutures, gomphoses, and syndesmoses (Table 3);

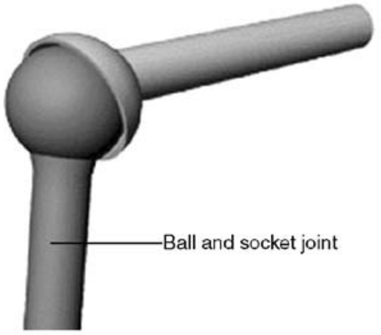
In addition to the obligatory components that all the synovial joints possess, several joints contain intraarticular structures. Discs and menisci are examples of such structures. They differ from one another mainly in that a disc is a circular structure that may completely subdivide a joint cavity so that it is, in reality, two joints in series, whereas a meniscus is usually a crescent-shaped structure that only partially subdivides the joint. Complete discs are found in the sternoclavicular and in the radiocarpal joint. A variety of functions have been proposed for intraarticular discs and menisci. They are normally met at locations where bone congruity is poor, and one of their main functions is to improve congruity and, therefore stability between articular surfaces. Shock absorption facilitation and combination of movements are among their likely roles. They may limit a movement or distribute the weight over a larger surface or facilitate synovial fluid circulation throughout the joint.

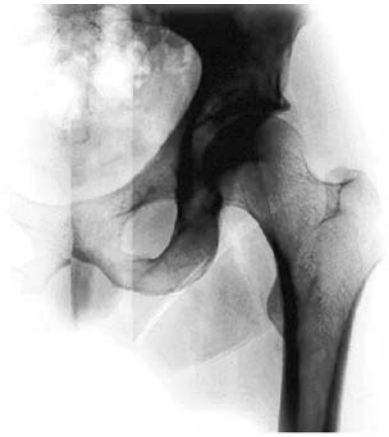
The labrum is another intraarticular structure. In humans, this structure is only found in the glenohumeral and hip joints. They are circumferential structures attached to the rim of the glenoid and acetabular sockets. Labra are distinct from articular cartilage because they consist of fibrocartilage and are triangular in their middle section. Their bases are attached to the articular margins and their free apical surfaces lined by synovial membrane. Like discs, their main function is to improve fit and protect the articular margins during extremes of movement.

Fat pads are localized accumulations of fat that are found around several synovial joints, although only those in the hip (acetabular pad) and the knee joint (infrapatellar pad) are named. Suggested functions for fat pads include protection of other intraarticular structures (e.g., the round ligament of the head of the femur) and serving as cushions or space-fillers thus facilitating more efficient movement throughout the entire available range.

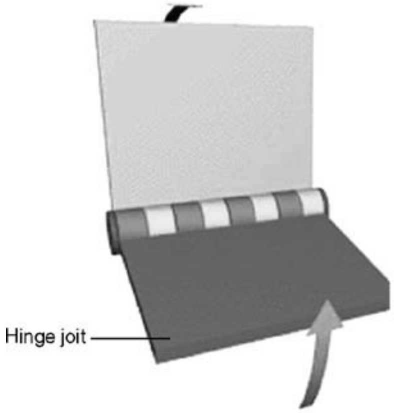
Bursae are enclosed, self-contained, flattened sacs typically with a synovial lining. They facilitate movement of musculoskeletal tissues over one another and thus are located between pairs of structures (e.g., between ligament and tendon, two ligaments, two tendons or skin, and bone). Deep bursae, such as the illiopsoas bursa or the deep

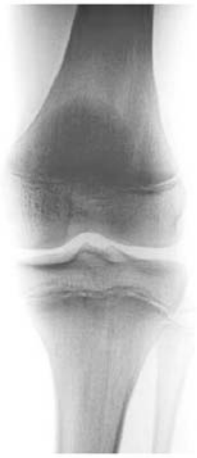
Table 1. Diarthroses: Synovial Joints



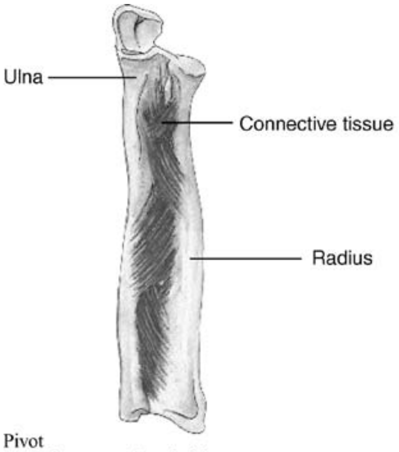


Ball and socket
Other names: Spheroidal; endarthroses
Description: Ball-shaped head fits into concave socket
Movement: Widest range of all joints; triaxial
Example: Shoulder and hip joints



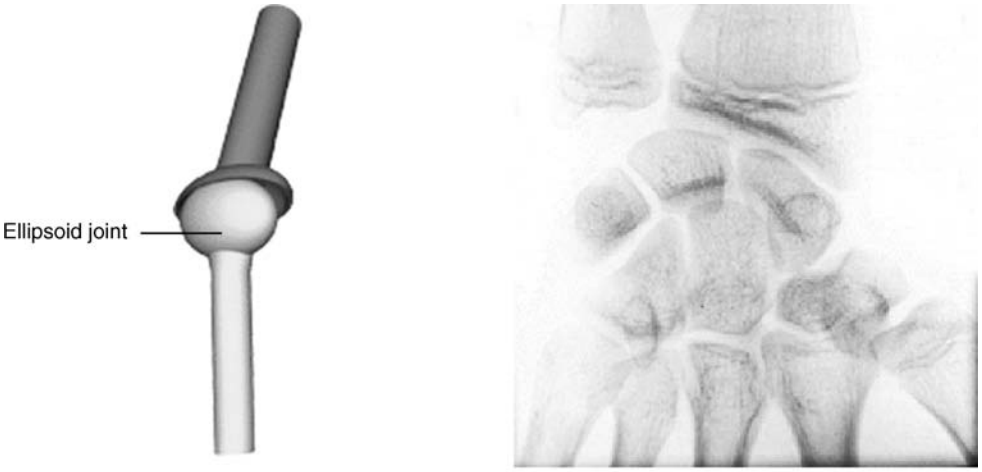


Hinge
Other name: Ginglymus
Description: Spool-shaped head fits into concave surface
Movement: In one plane about single axis (uniaxial); like hinged-door movement (namely, flexion and extension)
Examples: Elbow, knee, ankle, and interphalangeal joints

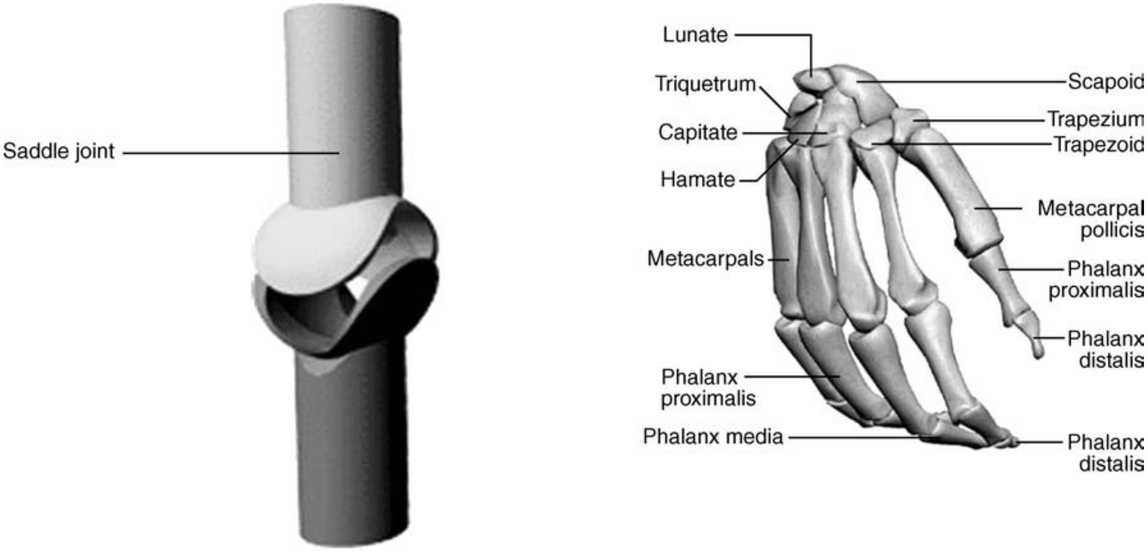


Pivot
Other name: Trochoid
Description: Arch-shaped surface rotates about rounded or peglike pivot
Movements: Rotation: uniaxial
Example: Between axis and atlas; between radius and ulna

Table 1. (*Continued*)



Ellipsoidal
- Other names: Condyloid, ovoid
- Description: Arch-shaped condyle fits into elliptical cavity
- Movements: In two planes at right angles to each other – specifically, flexion, extension, abduction, and adduction; biaxial
- Example: Between radius and carpals



Saddle
- Other name: Reciprocal
- Description: Saddle-shaped bone fits into socket that is concave-convex in opposite direction; modification of condyloid joint
- Movements: Same kinds of movement as condyloid joint but freer; like rider in saddle; biaxial
- Example: Thumb, between first metacarpal and trapezium

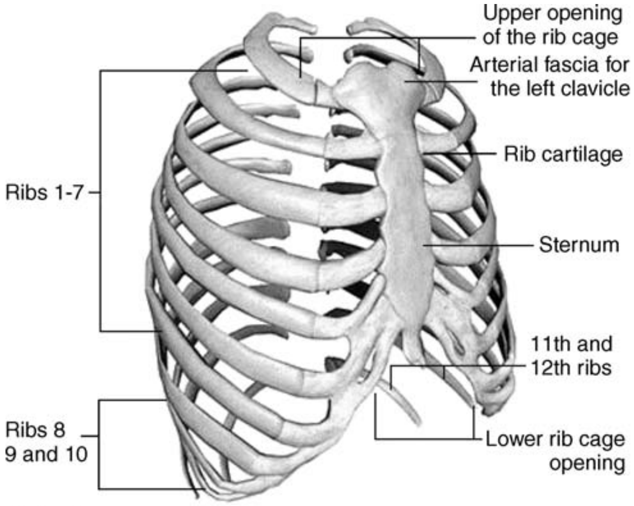
Gliding
- Other name: Arthroidal
- Description: Articulating surfaces; usually flat
- Movement: Gliding, a nonaxial movement
- Example: Between carpal bones; between sacrum and ilium (sacroiliac joints)

retrocalcaneal bursa, develop along with joints and by a similar series of events during the embryonic period.

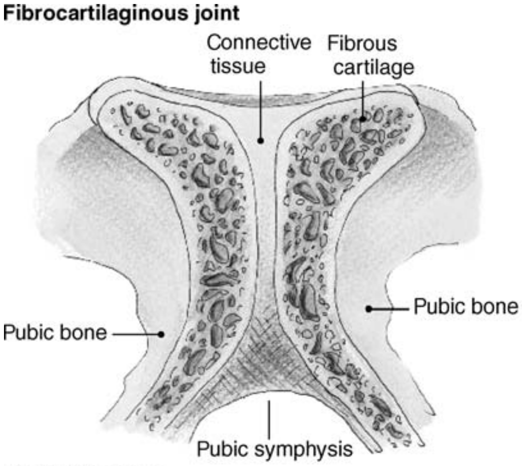
Tendons are located at the ends of many muscles and are the means by which these muscles are attached to bone or other skeletal elements. The primary structural component of tendons is type I collagen. Tendons almost exclusively operate under tensile forces.

Ligaments are dense bands of connective tissue that connect skeletal elements to each other, either creating (as in the case of syndesmoses) or supporting joints. According to their location they are classified as intracapsular, capsular, or extracapsular. Structurally, they resemble the tendons in that they consist predominantly of type I collagen.

Table 2. Amphiarthroses: Cartilaginous Joints



Cartilaginous
Other name: Synchondroses
Description: The joint formed by each of the costal cartilage
Movement: Bending and twisting, or slight compression
Example: Between the ribs and the sternum; between carpal and tarsal bones



Fibrocartilaginous
Other name: Symphyses
Description: Within the joint, separating the bones, is a fibrocartilaginous pad – these pads, or discs, serve as shock absorbers
Movement: Compression, flexion, extension, and rotation
Example: Intervertebral and pubic joints

Articular Cartilage

Articular cartilage, the resilient load-bearing tissue that forms the articulating surfaces of synovial joints functions through load distribution mechanism by increasing the area of contact (thereby reducing the stress) and provides these surfaces with the low friction, lubrication, and wear characteristics required for repetitive gliding motion.

Biomechanically, cartilage is another intraarticular absorption mechanism that dampens mechanical shocks and spreads the applied load onto subchondral bone (Fig. 2). Articular cartilage should be viewed as a multiphasic material. It consists primarily of a large extracellular matrix (ECM) with a sparse population of highly specialized cells (chondrocytes) distributed throughout the tissue. The primary components of the ECM are water, proteoglycans, and collagens, with other proteins and glycoproteins present in lower amounts (1). The solid phase is comprised by this porous-permeable collagen-PG matrix filed with freely movable interstitial fluid (fluid phase) (2). A third phase is the ion phase, necessary to describe the electromechanical behaviors of the system. The structure and composition of the articular cartilage vary throughout its depth (Fig. 2), from the articular surface to the subchondral bone. These differences include cell shape and volume, collagen fibril diameter and orientation, proteoglycan concentration, and water content. These all combine to provide the tissue with its unique and complex structure and mechanical properties. A fine mechanism of interstitial fluid pressurization

Table 3. Synarthroses: Fibrous Joints

Sutures
Other name: —
Description: The edges of the bones have interdigitations or grooves that fit very closely and firmly together; the connecting fibers are very short
Movement: None
Examples: Between the flat bones of the skull

Syndesmoses
Other name: Ligamentous
Description: Two bones, that may be widely separated tied together by ligaments; the ligaments may be in the form of cords, bands, or flat sheets
Movement: None (some give)
Example: Between the distal ends of the tibia and fibula

Gomphosis
Other name: —
Description: A joint in which the surfaces of bony components are adapted to each other like a peg in a hole
Movement: None
Example: The conical process of a tooth is inserted in the bony socket of the mandible or maxilla

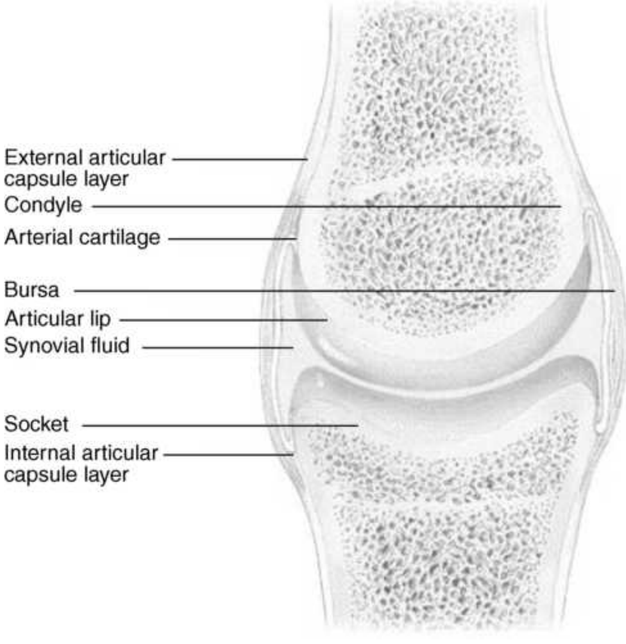

External articular capsule layer
Condyle
Arterial cartilage
Bursa
Articular lip
Synovial fluid
Socket
Internal articular capsule layer

Figure 1. Basic structure and components of a synovial joint (also called diarthroses).

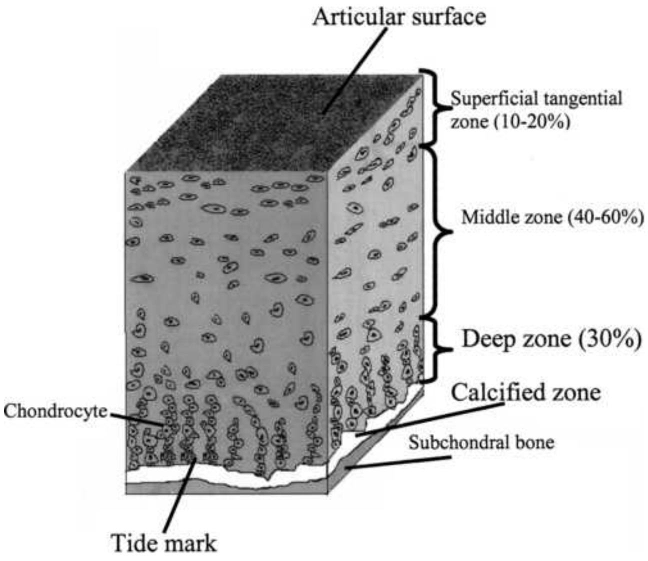


Figure 2. Zones of articular cartilage.

results from the flow of interstitial fluid through the porous-permeable solid matrix that in turn defines the rate dependent load-bearing response of the material. It is noteworthy that articular cartilage provides its essential biomechanical functions for eight decades or more in most of the human synovial joints and no synthetic material performs this well as a joint surface.

The frictional characteristics between two surfaces sliding over each other are significantly influenced by the topography of the given surfaces. Anatomical shape changes affect the way in which loads are transmitted across joints, altering the lubrication mode in that joint and, thus, the physiologic state of cartilage. Articular surfaces are relatively rough, compared to machined bearing surfaces, at the microscopic level. The natural surfaces are surprisingly much rougher than joint replacement prostheses. The mean of the surface roughness for articular cartilage ranges from 1 to 6 μm, while the metal femoral head of a typical artificial hip has a value of $\sim$0.025 μm, indicating that the femoral head is apparently much smoother. Topographic features on the joint surfaces are characterized normally by primary anatomic contours, secondary roughness ($<$0.5 mm in diameter and $<$50 μm deep), tertiary hollows on the order of 20–45 μm deep; and, finally, quaternary ridges 1–4 μm in diameter and 0.1–0.3 μm deep. Scanning electron micrographs (SEMs)of arthritic cartilage usually depict a large degree of surface irregularity and anomalous microtopography. These surface irregularities have profound effects on the lubrication mechanism. They accelerate the effects of friction and the rate of degradation of the articular cartilage. The types of joint surface interactions vary greatly between different joints in the body, different animals, between different size animals of the same species, different genders, and different ages. For example, the human hip joint is a deep congruent ball and socket joint (where the cartilage thickens peripherally at the acetabulum); this differs greatly from the bcondylar nature of the distal femur in the knee joint, and the saddle shape of the carpometacarpal joint in the thumb. The degree of shape matching between the various bones and articulating cartilage surfaces composing a joint is a major factor affecting the distribution of stresses in the cartilage and subchondral bone.

Effects of Motion and External Loading on Joints

The articular joint is viewed as an organ with complicated mechanisms of memory and adaptation that accommodates changes in its function. Joint loading results in motion and the couple load–motion is required to maintain normal adult articular cartilage composition, structure, and mechanical properties. The type, intensity, and frequency of loading necessary to maintain normal articular cartilage vary over a broad range. The intensity or frequency of loading should not exceed or fall below these necessary levels, since this will disturb the balance between the processes of synthesis and degradation. Changes in the composition and microstructure of cartilage will result. Reduced joint loading, as has been observed in cases of immobilization by casting or external fixation, leads to atrophy or degeneration of the cartilage. The changes affect both the contact and noncontact areas. Changes in the noncontact areas resulting from rigid immobilization include fibrillation, decreased proteoglycan content and synthesis, and altered proteoglycan conformation, such as a decrease in the size of aggregates and amount of aggregate. Normal nutritive transport to cartilage from the synovial fluid by means of diffusion and convection has been diminished, resulting in these changes. Increased joint loading, either through excessive use, increased magnitudes of loading, or impact, also may affect articular cartilage. Catabolic effects can be induced by a single-impact or repetitive trauma, and may serve as the initiating factor for progressive degenerative changes. Osteoarthritis, a joint disease of epidemic proportions in the western world, is characterized by erosive cartilage lesions, cartilage loss and destruction, subchondral bone sclerosis and cysts, and large osteophyte formation at the margins of the joint (3).

Moderate running exercise may increase the articular cartilage proteoglycan content and compressive stiffness, decrease the rate of fluid flux during loading, and increase the articular cartilage thickness in skeletally immature animals. However, no significant changes in articular cartilage mechanical properties were observed in dogs in response to lifelong increased activity that did not include high impact or torsional loading of their joints. Disruption of the intraarticular structures (e.g., menisci or ligaments) will alter the forces acting on the articular surface in both magnitude and areas of loading. The resulting joint instability is associated with profound and progressive changes in the biochemical composition and mechanical properties of articular cartilage. In experimental animal models, responses to transection of the anterior cruciate ligament or meniscectomy have included fibrillation of the cartilage surface, increased hydration, changes in the proteoglycan content, reduced number and size of proteoglycan aggregates, joint capsule thickening, and osteophyte formation. It seems likely that some of these changes result from the activities of the chondrocytes, because their rates of synthesis of matrix components, breakdown of matrix components, and secretion of proteolytic enzymes are all increased. *In vitro* studies have shown that loading of the cartilage matrix can cause all of these mechanical, electric, and physicochemical events, but thus far it has not been clearly demonstrated which signals are most important in stimulating the anabolic and catabolic activity of the chondrocytes. A holistic physicochemical and biomechanical model of cartilage function in health and disease remains a challenge in the scientific community.

KINEMATICS OF JOINTS

General Comments

Mechanical analysis can refer to kinetics (forces) and/or kinematics (movement), with kinetics being the cause and kinematics the result. Mechanical analysis can develop models proceeding from forces to movements or vice versa. The analysis that starts from the cause (force) is called direct or forward dynamics, and produces a defined set of forces that caused the unique movement. This approach has one solution, and hence is deterministic. Starting from the movement the analysis is called inverse dynamics. In this case, an infinite number of combinations of individual forces acting on the system can be the causes of the same unique movement, which makes the inverse dynamics approach not deterministic. The simplest and most essential system of mechanical formulations for explaining and describing motion is the Newton's second law. More advanced techniques include the Lagrange, d'Alembert, and Hamilton's methods. In general all of these methods start by describing equations of motion for a rigid body for translation, rotation, or combinations of them for both two (2D) and three-dimensional (3D) space. If the model assumes that the articulated segments that create the articulation are modeled as rigid bodies the remaining task is to calculate the relative motion between the two segments by applying graphics or joint kinematic analysis.

Kinematics is the study of the movements of rigid structures, independent of the forces that might be involved. Two types of movement, translation (linear displacement) and rotation (angular displacement), occur within three orthogonal planes; that is, movement has six degrees of freedom. Humans belong to the vertebrate portion of the phylum chordata, and as such possess a bony endoskeleton that includes a segmented spine and paired extremities. Each extremity is composed of articulated skeletal segments linked together by connective tissue elements and surrounded by skeletal muscle. Motion between skeletal segments occurs at joints. Most joint motion is minimally translational and primarily rotational. The deviation from absolute rotatory motion may be noted by the changes in the path of a joint's "instantaneous center of rotation". These paths have been measured for most of the joints in the body and vary only slightly from true arcs of rotation. For human motion to be effective, not only must a comparatively rigid segment rotate its position relative to an adjacent segment, but many adjacent limb movements must interact as well. Whether the hand is trying to write or the foot must be lifted high enough to clear an obstacle on the ground, the activity is achieved via coordinated movements of multiple limb segments. To provide for the greatest possible function of an extremity, the proximal joint must have the widest range of motion to position the limb in space. This joint must allow for rotatory motions of large degrees in all three planes about all three axes. A means is also provided to translate the limb, so that an extremity can function at all locations within its global range. Rotational motion of the elbow and knee joints allows such overall changes as adjacent limb segments move. Finally, to fine-tune the use of this mechanism with respect to the extremities, for their functional purposes, the hand and foot are required to have a vast amount of movement about all three axes, although the rigid segments are relatively small. Such movement requires the presence of relatively universal joints at the terminal aspect of each extremity.

Characterization of the General Mechanical Joint System: Terminology and Definitions

The displacement of a point is simply the difference between its position after a motion and its position before that motion. It can be represented by a 3D vector drawn from the initial position of the point to its final position. The components of the displacement vector will be the changes in the coordinates of the point's position from measurement in the reference coordinate system. It is apparent that not only the positions, but also displacements measured are relative to some reference. Rigid body (RB) displacements are more complicated than point displacements since for a rigid body a displacement is a change in its position relative to some reference, but more than three parameters are needed to describe it. Two simple types of RB displacement can be described: translation and rotation. An important property of pure translation of a RB is that the displacement vectors of all points in the body are identical and are nonzero. In pure rotation of a RB, although points in the body experience nonzero displacements, one point in that body experiences zero displacement. In addition to that rule, Euler's theorem shows that in pure rotation all points along a particular line through that undisplaced point also experience zero displacement. This line is also known as the axis of rotational displacement. Chasles theorem further states that any displacement of a RB can be accomplished by a translation along a line parallel to the axis of rotation that is defined by Euler's theorem plus a rotation about that same parallel axis. Simply that suggests that any displacement in 3D is equivalent to the motion of a nut, representing the body, on an appropriate stationary screw that was centered on the line described above. Indeed, it can be shown that any displacement in 3D is equivalent to a translation plus a rotation.

Degrees of Freedom

The biological organisms capable of propelling themselves through different media consist of more than one rigid body. A system consisting of a 3D reference frame and an isolated rigid body in space has six degrees of freedom (DOF). To describe the position of each body relative to the ground reference frame it would be necessary to use six parameters, so for two unconnected rigid bodies 12 parameters would be necessary. The system consisting of these two unconnected bodies and the fixed -ground reference would have 12 DOF. The human–animal body consists of a combination of suitably connected bodies. The connections, joints between the bodies, serve to constrain the motions of the bodies so that they are not free to move with what would otherwise be six DOF for each body. Therefore, we can define the number of DOF that the joint removes as the number of degrees of constraint that it provides. It can be shown that every time a joint is added to a system, the number of degrees of freedom in that system is reduced by the number of degrees of constraint provided by that joint. This suggests the following generic formula for the calculation of the degrees of freedom of a system:

$$F = 6(L-1) - 5J_1 - 4J_2 - 3J_3 - 2J_4 - J_5$$

where $F =$ is the number of degrees of freedom in the system of connected joints; $L =$ is the number of joints in the system, including the ground joint (which has no degrees of freedom), and $J_n =$ the number of joints having n degrees of freedom each.

Table 4 contains a description of the major joints in the human body along with the segments-bones that they

Table 4. Characteristics of Major Human Joints

Joint	Bones	Type	DOF	Type of motion	Range of Motion (deg)
Shoulder	Humerus-Scapula	Diarthrosis (spheroidal)	3	Flexion	150
				Extension	50–60
				Abduction	90–120
				Abduction	Complete
				Rotation	
Elbow	Humerus-ulna	Diarthrosis (ginglymus)	2	Flexion	145–160
				Extension	0–5
				Rotation (radius)	
Radioulnar	Superior radius-ulna	Diarthrosis (trochoid)	1	Pronation	70–75
				Supination	85–90
Wrist	Radius-carpal	Diarthrosis (condyloid)	2	Flexion	90–95
				Extension	60–70
				Radial deviation	20–25
				Ulnar deviation	55–65
				Circumduction	Complete
Metacarpal-phalangeal	Metacarpal-phalanges	Diarthrosis (condyloid)	2	Flexion	80–90
				Extension	20–30
				Radial deviation	20–25
				Ulnar deviation	15–20
Finger	Interphalanges	Diarthrosis (ginglymus)	1	Flexion	80–90
				Extension	0–10
Thumb	First metacarpal-carpal	Diarthrosis (reciprocal)	2	Flexion	80–90
				Extension	20–25
				Abduction	40–45
				Abduction	0–10
				Circumduction	Complete
Hip	Femur-acetabulum	Diarthrosis (spheroidal)	3	Flexion	90–120
				Extension	10–20
				Abduction	30–45
				Abduction	30
				Medial rotation	30–40
				Lateral rotation	60
				Circumduction	Complete
Knee	Tibia-femur	Diarthrosis (ginglymus)	2	Flexion	120–140
				Extension	0
				Medial rotation	30
				Lateral rotation	40
Ankle	Tibia-fibula-talus	Diarthrosis (ginglymus)	1	Flexion	20–30
				Extension	40–45
Intertarsal	Tarsals	Diarthrosis (arthroidal)	2	Gliding	Limited motion
Metatarsal-phalangeal	Metatarsals-phalanges	Diarthrosis (condyloid)	2	Flexion	25–30
				Extension	80–90
				Abduction	15–20
				Abduction	Limited
Interphalangeal	Phalanges	Diarthrosis (arthroidal)	1	Flexion	90
				Extension	0
Tibio-fibular	Distal tibia-fibula	Synarthrosis (syndesmosis)	0	Slight movement	Give
Skull	Cranial	Synarthrosis (suture)	0	No movement	
Sterno-costal	Ribs-sternum	Amphiarthrosis (synchondrosis)	0	Slight movement	
Sacroiliac	Sacrum-ilium	Amphiarthrosis (synchondrosis)	0	No movement	Elastic
Intervertebral	Cervical vertebrae	Diarthrosis (arthroidal)	3	Flexion	40
				Extension	75
				Lateral flexion	35–45
				Axial rotation	45–50
	Thoracic vertebrae	Diarthrosis (arthroidal)	3	Flexion	105
				Extension	60
				Lateral flexion	20
				Axial rotation	35
	Lumbar vertebrae	Diarthrosis (arthroidal)	3	Flexion	60
				Extension	35
				Lateral flexion	20
				Axial rotation	5
	Atlas axis	Diarthrosis (trochoid)	1	Pivoting motion	

articulate, their respective type DOF and type/range of motion they provide.

Planar Motion

Some human joints move predominantly in one plane (e.g., the knee joint) in which case the motion can be approximated and analyzed by graphical methods. Here the rotation is characterized by the motion of all points on concentric cycles with an identical angle of rotation around the undisplaced center of rotation (CR). The CR may be located inside and outside of the boundaries of the rotating body. The most common graphical method for the calculation is the so-called bisection technique. If the initial and final states of the body are known, the position of the center or rotation and the angle of rotation may be reconstructed (Fig. 3).

Instantaneous Center of Rotation

When a 2D body is rotating without translation, for example, a rotating stationary bicycle gear, any marked point P on the body may be observed to move in a circle about a fixed point called the axis of rotation or center of rotation. When a rigid body is both rotating and translating, for example, the motion of the femur during gait, its motion at any instant of time, can be described as rotation around a moving center of rotation. The location of this point at any instant, termed the instantaneous center of rotation (ICR), is determined by finding the point which, at that instant, is not translating. Then by definition, at that instant, all points on the rigid body are rotating about the ICR. For practical purposes, the ICR is determined by noting the paths traveled by two points, S and Q, on the object in a very short period of time to S$'$ and Q$'$. The paths SS$'$ and QQ$'$ will be perpendicular to lines connecting them to the ICR because they approximate, over short periods, tangents to the circles describing the rotation of the body around the ICR at that instant. Perpendicular bisectors to these two paths will intersect at the instantaneous axis of rotation.

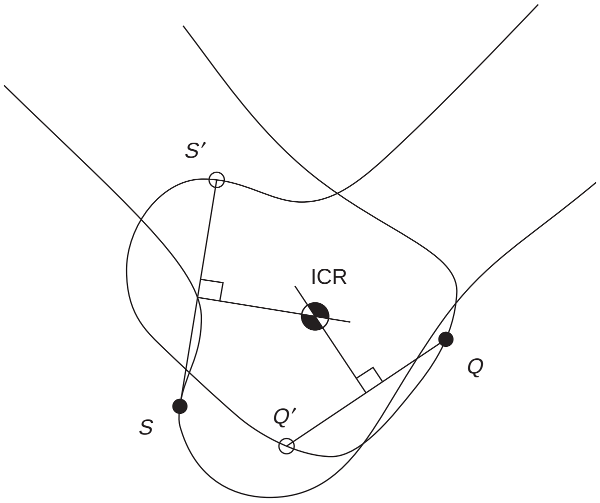


Figure 3. Points S and S′ as well as Q and Q′, lie on the arcs of circles around the center of rotation ICR (used synonymously with CR after the section Instantaneous Center of Rotation). If lines SS′ and QQ′ are bisected perpendicularly, the center of rotation CR is located at the intersection of these perpendicular bisectors. This construction assumes that the perpendicular bisectors are differently orientated but a special case arises if the bisectors are identically oriented. Then the points S, Q and the center of rotation ICR lie on a straight line.

If the ICR is considered to be a point on a moving body, its path on the fixed body is called a fixed centrode. If the ICR is considered to be a point on a fixed body, its path on the moving body is called the moving centrode.

Although in principle two objects may move relative to one another in any combination of rotation and translation, diarthroidal joint surfaces are constrained in their relative motion. The articular surface geometry, the ligamentous restraints, and the action of muscles spanning the joint are the main constraining systems. In general, joint surface separation (or gapping-proximal-distal) and impaction are small compared to overall joint motion. Mechanically, when surfaces are adjacent to each other they may move relative to each other in either sliding or rolling contact. In rolling contact, the contacting points on the two surfaces have zero relative velocity, that is, no slip. Rolling and sliding contacts occur together when the relative velocity at the contact point is not zero. The instant center will then lie between the geometric center and the contact point. All diarthroidal joint motion consists of both rolling and sliding motion. In the hip and shoulder, sliding motion predominates over rolling motion. In the knee, both rolling and sliding articulation occur simultaneously. These simple concepts affect the design of total joint prostheses. For example, some total knee replacements have been designed for implantation while preserving the posterior cruciate ligament, which appears to help maintain the normal kinematics of rolling and sliding in the knee. Other knee prostheses substitute for ligament control of kinematics by alterations in articular surface contour through constraining congruity.

Analytical Methods

Simple kinematic analysis of pure planar translations and rotation or combinations of the two as well as complicated 3D analysis of a rigid body requires the positional information of a minimum of three noncolinear points to describe this motion uniquely. If the position of three points at two instants is known, the displacement from one position to another may be interpreted as translation, rotation or both. Therefore, the first task is to continually monitor the positions of three points on each rigid body. This analysis is conveniently divided into data collection and data analysis.

Data Collection. A constant challenge for the experimental motion analyst is the collection of accurate spatial displacement kinematics of a joint. Several methods have been employed. A review is presented here.

Video and digital optical motion capture (tracking) systems offer state-of-the-art, high resolution, accurate motion capture options to acquire, analyze, and display 3D motion data. The systems are integrated with analog data acquisition systems to enable simultaneous acquisition (1–300 Hz) of force plate and electromyographic data. Clinically validated software analysis packages are used to analyze and display kinematic, kinetic, and electromyographic data in forms that are easy to interpret.

The major components of a video and digital motion capture system are the cameras, the controlling hardware modules, the software to analyze and present the data, and the host computer to run the software. These systems are designed to be flexible, expandable (from 3 to up to 200 cameras in motion analysis tracking for Hollywood animation movies) and easy to integrate into any working environment. This system collects and processes coordinate data in the least amount of time and requires minimal operator intervention. This system uses motion capture cameras to rapidly acquire 3D coordinate positions from reflective markers placed on subjects. Illuminating strobes with differing wavelengths are used to track the spatial displacement (between 1 and 10 mm resolution) of spherical reflective markers attached to the subject's skin at appropriately chosen locations, preferably on bony landmarks on the human body to minimize skin movement. They can be infrared (IR), visible red, or near-IR strobes to fit the lighting conditions of the capture environment. Also, the lenses can be of fixed or variable focal length for total adaptability. Images are processed within the optical capture cameras where markers are identified and coordinates are calculated before being transferred to the computers. After the completion of the movement, the system provides 3D coordinate and kinematic data. The disadvantages of the system include the skin movement error whose effect is more prominent (3 cm error) at high movement speeds. These high speed motion tasks (impact biomechanics, e.g.) are handled by high speed cine cameras with data acquisition rates several orders of magnitude greater than clinical motion analysis systems. The processing method that is almost real time uses combinations of skin markers (minimum three at each segment) to produce coordinate systems for each segment and eventually describe intersegmental relative motion or relate all the different segment motions to the laboratory fixed-coordinate system. Recently, methods employing clusters of markers have shown to somewhat reduce the skin marker artifact but are yet to be adopted in the clinical practice.

A more accurate method (<1 mm translation and up to 1000 Hz) is the cineradiographical method, which employs an X-ray machine and uses special cameras for capture of sequences of the digital radiographs. In addition to accuracy, these systems directly access the *in vivo* skeletal kinematics so that the resulting analysis can be directly related to bony landmarks. Radiation issues, magnification, and distortion factors are some drawbacks that can be overcome by appropriate image analysis techniques. This method is, however, prone to occlusion errors when two segments overlap and simultaneously cross the field of view of the X-ray source. Stereosystems with more than one X-ray sources can limit this artifact. A biplane radiographic system consists of two X-ray generators and two image intensifiers optically coupled to synchronized high speed video cameras that can be configured in a custom gantry to enable a variety of motion studies. The system can be set up with various set-up modes (e.g., a 60° interbeam angle), an X-ray source to object distance of 1.3 m, and an object to intensifier distance of 0.5 m. Images are acquired with the generators in continuous radiographic mode (typically 100 mA, 90 kVp). The video cameras are electronically shuttered to reduce motion blur. Short (0.5 s) sequences are recorded to minimize radiation exposure. X-ray exposure and image acquisition are controlled by an electronic timer–sequencer to capture only the desired phase of movement.

CODA is an acronym of Cartesian Opto-electronic Dynamic Anthropometer, a name first coined in 1974 to give a working title to an early research instrument developed at Loughborough University, United Kingdom. The 3D capability is an intrinsic characteristic of the design of the sensor units, equivalent to but much more accurate than the stereoscopic depth perception in normal human vision. The system is precalibrated for 3D measurement, which means that the lightweight sensor can be set up at a new location in a matter of minutes, without the need to recalibrate using a space-frame. Each sensor unit must be independently capable of measuring the 3D coordinates of skin markers in real-time. As a consequence, there is great flexibility in the way the system can be operated. For example, a single sensor unit can be used to acquire 3D data in a wide variety of projects, such as unilateral gait. Up to six sensor units can be used together and placed around a capture volume to give "extra sets of eyes" and maximum redundancy of viewpoint. This enables the system to track 360° movements that often occur in animation and sports applications. The calculation of the 3D coordinates of markers is done in real-time with an extremely low delay of 5 ms. Special versions of the system are available with latency shorter than 1 ms. This opens up many applications that require real-time feedback, such as research in neurophysiology and high quality virtual reality systems, as well as tightly coupled real-time animation. It is also possible to trigger external equipment using the real-time data. The automatic intrinsic identification of markers combined with processing of all 3D coordinates in real-time means that graphs and stick figures of the motion and many types of calculated data can be displayed on a computer screen during and immediately after the movement occurs. The data are also immediately stored to file on the hard drive.

A new concept in measuring movement disorders utilizes a unique miniature solid-state gyroscope, not to be confused with gravity sensitive accelerometers. The instrument is fixed with straps directly on the skin surface of the structure whose motion is of interest. It has been successfully used to quantify: tremor (resting, posture, kinetic), rapid pronation–supination of the hand, arm swing, lateral truncal sway, leg stride, spasticity (pendulum drop test), dyskinesia, and alternating dystonia. The system (Motus) senses rotational motion only and is ideal for quantifying human movement since most skeletal joints produce rotational motion. This disadvantage is outweighed by its miniature size that allows it to be of great value for certain types of studies. A different system (Gypsy Gyro) uses 18 small solid-state inertial sensors (gyros) to accurately measure the exact rotations of the actor's bones in real-time for motion capture. The system can easily be worn beneath normal clothing. With wireless range these systems–suits can be used to record up to 64 actors simultaneously.

Another concept for 3D motion analysis is the measurement system CMS10 (Zebris) designed as a compact device for everyday use. The measurement procedure is based on the travel time measurement of ultrasonic pulses that are emitted by miniature transmitters (markers placed on the skin) to the three microphones built into the compact device. A socket for the power pack (supplied with the device) as well as the interface to a computer are located on the back of the device. The evaluation and display of the measurement data are carried out in real-time. It is possible to use either a table clamp or a mobile floor stand with two joints to support the measurement system.

Data Analysis

Coordinate Systems and Transformation. In the analysis of experimental joint mechanics data, the transformation of point coordinates from one coordinate system to another is a frequent task (4). A typical application of such a transformation would be gait analysis data recorded in a laboratory fixed coordinate system (by means of film or video sequences) that must be converted to a reference system fixed to the skeleton of the test subject. The laboratory fixed coordinate system may be designated by xyz and the body reference system by abc (Fig. 4). The location of a point $S(a/b/c)$ in the body reference system is defined by the radius vector $s = a \cdot e_a + b \cdot e_b + c \cdot e_c$. Consider the reference system to be embedded into the laboratory system. Then the radius vector $r_m = x_m \cdot e_x + y_m \cdot e_y + z_m \cdot e_z$ describes the origin of the reference system in the laboratory system. The location of $S(x/y/z)$ is now expressed by the coordinates a, b, c. The vector equation $r = r_m + s$ gives the radius vector for point S in the laboratory system (Fig. 4). Employing the full notation we have: $r = (x \cdot e_x + y \cdot e_y + z \cdot e_z) = (x_m \cdot e_x + y_m \cdot e_y + z_m \cdot e_z) + (a \cdot e_a + b \cdot e_b + c \cdot e_c)$. A set of transformation equations results after some intermediate matrix algebra to describe the coordinates. The scalar products of the unit vectors in the xyz and abc systems produce a set of nine coefficients C_{ij}. The cosine of the angle between the coordinate axes of the two systems corresponds to the value of the scalar products. Three "direction cosines" define the orientation of each unit vector in one system with respect to the three unit vectors of the other system. Due to the inherent properties of orthogonality and unit length of the unit vectors, there are six constraints on the nine direction cosines, which leaves only three independent parameters describing the transformation. Employing the matrix notation of the transformation equation we have

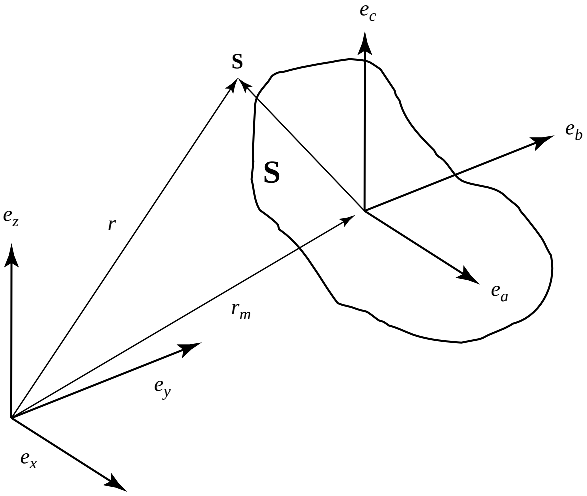


Figure 4. Changing the coordinate systems, transformation of point coordinates from one coordinate system to another.

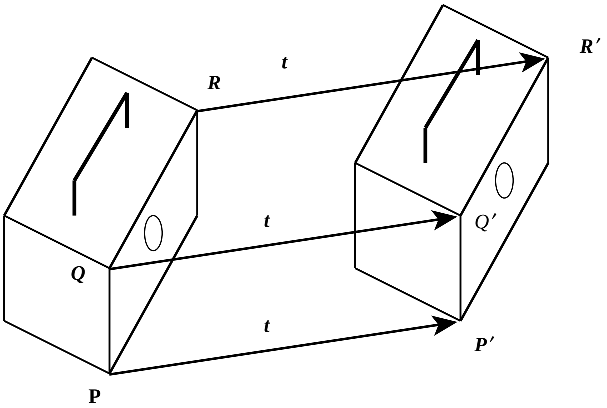


Figure 5. A rigid body (shoebox) moves parallel to itself. The radius vectors from O to P and from O to P' are designated by r and r' so that $r' = r + t$, where t is the difference vector.

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} x_m \\ y_m \\ z_m \end{bmatrix} + \begin{bmatrix} c_{11}\, c_{12}\, c_{13} \\ c_{21}\, c_{22}\, c_{23} \\ c_{31}\, c_{32}\, c_{33} \end{bmatrix} * \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

In coordinate transformations, the objects remain unchanged and only their location and orientation are described in a rotated and possibly translated coordinate system. If a measurement provides the relative spatial location and orientation of two-coordinate systems the relative translation of the two systems and the nine coefficients C_{ij} can be calculated. The coefficients are adequate to describe the relative rotation between the two coordinate systems.

Translation in Three-Dimensional Space. In translation in 3D space, the rigid object moves parallel to itself (Fig. 5). Pure translation in 3D space leaves the orientation of the body unchanged as in the case of pure 2D translation.

Rotations about the Coordinate Axes. A rotation in 3D space is defined by specifying an axis and an angle of rotation (Fig. 6). The axis can be described by its 3D orientation and location (5). A rotation, as does the translation explained earlier, leaves all the points on the axis unchanged; all other points move along circular arcs in planes oriented perpendicular to the axis (6,7).

This rotation moves an arbitrary point P to location P' with constant distance z from the xy plane ($z = z'$). This produces the following matrix notation for the respective equations for the rotation that changes x and y coordinates

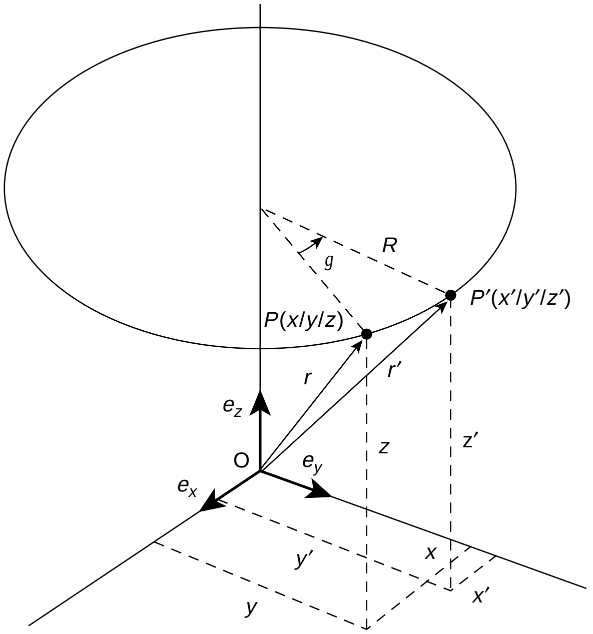


Figure 6. Rotation about the z axis of the coordinate system.

but leaves the z coordinate unchanged.

$$r' = \begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} \cos\gamma & -\sin\gamma & 0 \\ \sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = D_z(\gamma)r$$

The matrix describing a rotation about the z axis is designated $\boldsymbol{D}_z(\lambda)$. The matrices describing a rotation about the y axis through angle β and about x axis through angle α are similar.

$$r' = \begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} \cos\beta & 0 & \sin\beta \\ 0 & 1 & 0 \\ -\sin\beta & 0 & \cos\beta \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix} = D_y(\beta)r$$

$$r' = \begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha \\ 0 & \sin\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = D_y(\alpha)r$$

Combined Rotations as a Result of a Sequence of Rotations. Assume that the first rotation of a rigid body occurs about the z axis of a coordinate system. The rotation matrix related to the unit vectors e_x, e_y, e_z, is

$$D_z(\gamma = 90^\circ) = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The second rotation occurs supposedly about the x' axis, that is, about a body-fixed axis on the body (previously rotated about its z axis). The rotation matrix related to the unit vectors e'_x, e'_y, e'_z, is

$$D_{x'}(\alpha = 90^\circ) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix}$$

Matrix intermediate calculation here gives

$$r'' = D_{z'} * D_{x'} * r$$

In this calculation the sequence of the matrices is very important especially as this sequence differs from what one might expect. First, the matrix of the second partial rotation acts on the vector $\boldsymbol{r}$ and then, in a second step on the matrix of the first partial rotation. If the sequence of the two partial rotations is interchanged, the combined rotation is described by

$$r'' = D_x * D_{z'} * r$$

For rotations about body-fixed axes it is true that in general, the matrix of the last rotation in the sequence of rotations is the first one to be multiplied by the vector to be rotated. The matrix $\boldsymbol{B}$ describing the image resulting from n partial rotations about body-fixed axes is composed according to the formula:

$$B_{\text{body - fixed}} = D_1 * D_2 * D_3 * \ldots * D_{n-1} * D_n$$

where the indexes indicate the sequence of the rotations. Alternatively, if the n rotation were to be produced about axes fixed in space (i.e., fixed in the ground, laboratory frame) and not about body-fixed axes, the sequence of the matrexes in the matrix product would be different:

$$B_{\text{space - fixed}} = D_n * D_{n-1} * \ldots * D_2 * D_1$$

Euler and Bryant–Cardan Angles. Any desired orientation of a body can be obtained by performing rotations about three axes in sequence. There are, however, many ways of performing three such rotations. One can do this task at random, but for reasons of clarity two conventions are frequently used: the Euler's and Bryant–Cardan's rotations. In the Euler notation, the general rotation is decomposed of three rotations about body-fixed axes in the following manner:

Rotation 1: about the z axis through the angle φ rotation matrix $D_z(\varphi)$ (Fig. 7).

Rotation 2: about the x' axis through the angle θ rotation matrix $D_{x'}(\theta)$.

Rotation 3: about the z'' axis through the angle ψ rotation matrix $D_{z''}(\psi)$.

The matrix describing Euler's combined rotation is given by the matrix product

$$B = D_z(\varphi) * D_{x'}(\theta) * D_{z''}(\psi)\ (\text{Euler})$$

According to the Bryant and Cardan the general rotation is decomposed of three rotations about body-fixed axes in the following manner:

Rotation 1: about the x axis through the angle φ_1 rotation matrix $D_x(\varphi_1)$ (Fig. 7).

Rotation 2: about the y' axis through the angle φ_2 rotation matrix $D_{y'}(\varphi_2)$.

Rotation 3: about the z'' axis through the angle φ_3 rotation matrix $D_{z''}(\varphi_3)$,

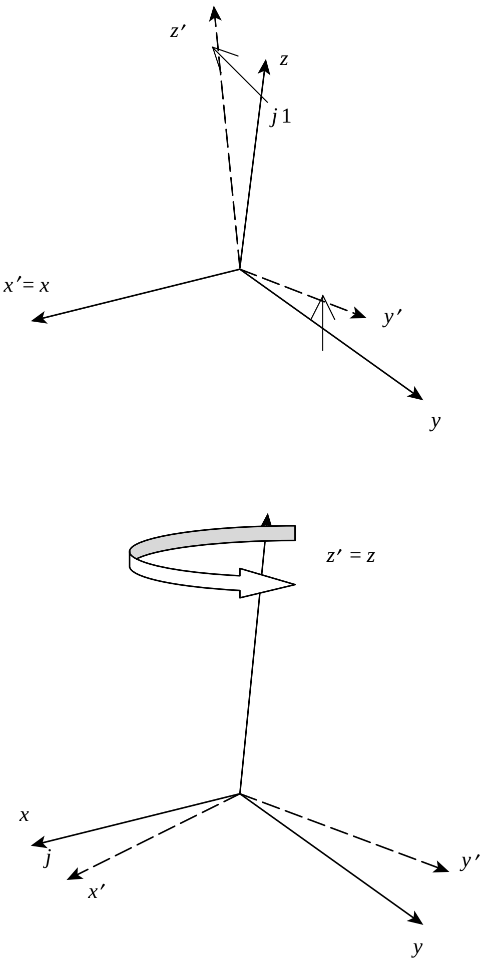


Figure 7. General rotation composed of three partial rotations. The first rotation according to the Bryant–Cardan convention (above). The first of the general rotations using Euler as the selection of the axes and angles of rotation (below).

in which case the matrix of combined rotation is given by

$$B = D_x(\varphi_1) * D_{y'}(\varphi_2) * D_{z''}(\varphi_3)\ (\text{Bryant} - \text{Cardan})$$

For reasons of simplicity, we have presented single or combined rotations about coordinate axes, but more complicated rotational laws can be applied as we deal with rotations about arbitrary axes. Rotation and translation can also be integrated into one single motion with Chasles theorem. Chasles theorem states that "the general motion in 3D space is helical motion", or "the basic type of motion adapted to describe any change of location and orientation in 3D space is helical motion". The relevant axis of rotation is designated the "helical axis". Chasles theorem is also known as the "helical axis" theorem.

KINETICS OF JOINTS

The study of the forces that bring about the movements discussed above is called kinetics. Because kinetics provides insights into the cause of the observed motion, it is essential to the proper interpretation of human movement processes. Forces and loads are not visually observable; they must be either measured with instrumentation or calculated from kinematics data. Kinetic quantities studied include such parameters as the forces produced by muscles; reaction loads between body parts as well as their interactions with external surfaces; the load transmitted through the joints; the power transferred between body segments; and the mechanical energy of body segments. Inherent to such studies are the functional demands imposed on the body. The structure and stability of each extremity and its joints reflect different systems and functional demands. The functional demands on the upper extremity are quite different from those on either the upper and lower axial skeleton or those on the lower extremity. Depending on which joint and/or structure is addressed, different types and degrees of rotational motion are allowed and are functional. How much structural strength is needed versus how much movement is allowed in each area dictates the nature of the material, size, shape, and infrastructure of the joint system established to perform a given movement.

Equations of Motion

The kinetics deal with the effects of forces on the motion of a body. When the motion is known, the problem is then to find the force system acting on the body. There are joint forces and joint moments. With all the kinematic quantities known, it is possible to find the joint forces and moments from the resulting force system that acts on each element. This is done by solving a system of simultaneous equations at successive time intervals. Since muscles are an unknown force system, the resolved muscle force and the real joint force are treated as totally unknown joint forces in the analysis. The three equations of motion for linear motions are

$$\sum F = ma_x \quad \sum F = ma_y \quad \sum F = ma_z$$

The three equations of motion for rotation are

$$\begin{aligned}\sum M_x &= I_{xx}\alpha_x - (I_{yy} - I_{zz})\omega_y\omega_z - I_{xy}(\alpha_y - \omega_x\omega_z)\\ &\quad - I_{yz}(\omega_y^2 - \omega_z^2) - I_{xx}(\alpha_z + \omega_x\omega_y)\\ \sum M_y &= I_{yy}\alpha_y - (I_{zz} - I_{xx})\omega_z\omega_x - I_{yz}(\alpha_z - \omega_y\omega_x)\\ &\quad - I_{xx}(\omega_z^2 - \omega_x^2) - I_{xy}(\alpha_x + \omega_y\omega_z)\\ \sum M_z &= I_{zz}\alpha_z - (I_{xx} - I_{yy})\omega_x\omega_y - I_{zx}(\alpha_x - \omega_z\omega_y)\\ &\quad - I_{xy}(\omega_x^2 - \omega_y^2) - I_{yz}(\alpha_y + \omega_z\omega_x)\end{aligned}$$

where M is the moment, I is the mass moment of inertia, α the angular acceleration, and ω is the angular velocity. The moment equations can be simplified if the axes of the reference frames coincide with the principal axes, with the origin at the center of gravity. These equations, called Euler equations, are

$$\begin{aligned}\sum M_x &= I_x\alpha_x - (I_y - I_z)\omega_y\omega_z\\ \sum M_y &= I_y\alpha_y - (I_z - I_x)\omega_z\omega_x\\ \sum M_z &= I_z\alpha_z - (I_x - I_y)\omega_x\omega_y\end{aligned}$$

Continuity conditions are derived based on the fact that equal and opposite forces and moments occur at the joint between the two segments.

The anthropometric data for the mass, the center of gravity, the moment of inertia, and so on for the different parts of the human body are available in the literature (8,9).

Motion and Forces on Diarthroidal Joints

In vivo experimental measurements on the relative motions between articulating surfaces of a joint, which correspond to daily activities, are limited. Most quantitative information is obtained from gait studies that do not provide the accuracy and precision for the detailed information required for lubrication studies. However, even simple calculations show that translational speeds between two articulating surfaces can range from $\sim 0.06\,\mathrm{m.s}^{-1}$ between the femoral head surface and the acetabulum surface during normal walking, to $\sim 0.6\,\mathrm{m.s}^{-1}$ between the humeral head surface and the glenoid surface of the shoulder when a baseball pitcher throws a fastball. Cartilage to cartilage contact or fluid-film layers, or a mixture of both are normally the contact mechanisms at the joint. During a normal walking cycle, the human hip, knee, and ankle joints can be subjected to loads on the order of six times body weight, with these peak loads occurring just after heel-strike and just before toe-off. The average load on the joint is approximately three to five times body weight, which lasts as long as 60% of the walking cycle. During the swing phase of walking, only light loads are carried. During this phase, the articular surfaces move rapidly over each other. In addition, extremely high forces occur across the joints in the leg during jumping. Descending stairs can load the knee with up to 10 times body weight, suggesting that the load on the joint surface is dependent on the task performed, that is, the loading sites change drastically as the articulating surfaces move relative to each other.

MATHEMATICAL AND MECHANICAL MODELS OF JOINTS

Locomotion results from complex, high dimensional, nonlinear, dynamically coupled interactions between an organism and its environment. Simple models called templates have been and can be made to resolve the redundancy of multiple legs, joints, and muscles by seeking synergies and symmetries. The simplest model (least number of variables and parameters) that exhibits a targeted behavior is called a template (10). Templates can be used to test control strategies against empirical data. Templates must be based in more detailed morphological and physiological models to ask specific questions about multiple legs, the joint torques that actuate them, the recruitment of muscles that produce those torques and the neural networks that activate the ensemble. These more elaborate models are called anchors. They introduce representations of specific biological details of the organism. The control of slow, variable-frequency locomotion appears to be dominated by the nervous system, whereas during rapid, rhythmic locomotion, the control may reside more within the mechanical system. Anchored templates of many-legged, sprawled-postured animals may suggest that passive, dynamic self-stabilization from a feedforward, tuned mechanical system can reject rapid perturbations and simplify control. Future progress would benefit from the creation of a field embracing comparative neuromechanics. Both templates and anchors are part of a system of mathematical and structural definitions and standard methods of description and dissemination of knowledge. In the next few sessions an attempt is made to describe some of those methods.

The human musculoskeletal system is often modeled by joining rigid links with continuous mass distribution. The joint may be of the revolute or spherical type, with restrictions consistent with body construction. Usually, the human segments form an open linkage.

Knowing all the kinematics at the center of mass of the segments, the joint force and the moment analysis proceeds by drawing free-body diagrams of the segments involved. The free body diagrams for the hip joint, the knee joint, and the ankle joint in a sagittal plane are illustrated in Fig. 8 as an example:

Assessment of Mechanical Factors Associated With Joint Degeneration: Limitations and Future Work

Joint degeneration results from complex, multidimensional, nonlinear, dynamically coupled interactions between the organism and its environment. The assessment of mechanical factors associated with joint degeneration has traditionally combined longitudinal clinical studies with carefully designed experimental techniques and theoretical computational analyses. The quality of

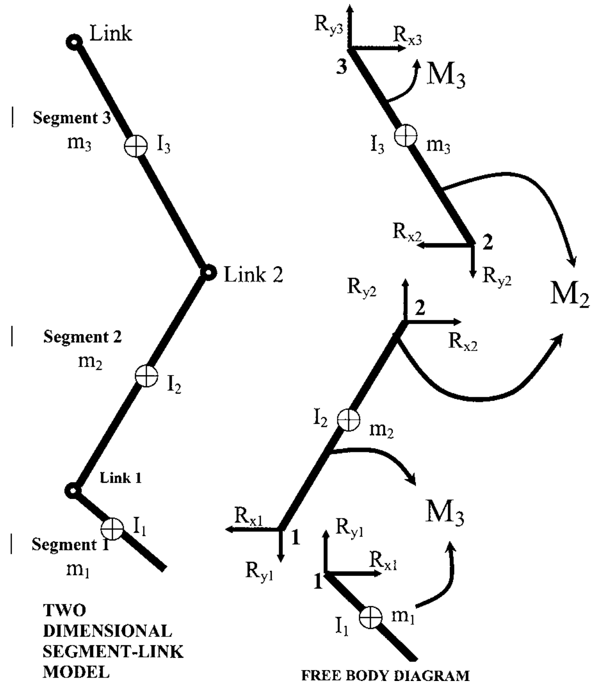


Figure 8. Relationship between the free body diagram and the link-segment model. Each segment is “broken” at the joints, and the reaction forces and moments of force acting at each joint are shown.

such assessments depends both on the accuracy–precision of the measurement methodology and the theoretical framework for its interpretation (i.e. joint mathematical models). To capitalize on the increasing level of measurement accuracy, theoretical analysis requires more detailed morphological and physiological models (11). For example, because of variations between individuals, the detailed, morphological analysis required for accurate modeling of cartilage stresses must be patient specific. In addition, the dynamics of the human task to be modeled (for e.g., human jumping) as expressed in the strain rate of tissue deformation must be accounted for in the analysis. Once the error estimate (simulation versus experiment) is established, model predictions can address specific clinical-biological questions. Traditionally, *in situ* methodology (cadaveric experiments and *in vitro* tests) is applied when an *in vivo* measurement is impossible. This limitation presents a number of implications and assumptions that weaken the theoretical analysis. Recent developments have further improved accuracy in the experimental measurement of *in vivo* knee kinematics. These developments allow significant improvements upon previous limitations by applying patient-specific task-dependent models in the study of joint pathogenesis.

The vast majority of dynamic knee studies have been performed with conventional motion analysis techniques, using markers attached to the skin. Conventional motion analysis is not sufficiently accurate to enable analysis of cartilage stress. Previous studies have shown that skin markers move substantially relative to underlying bone, with RMS errors of 2–7 mm and peak errors as large as 14 mm in estimates of tibial position during gait (12). A study of four subjects during running and hopping (using 250 frame$\cdot$s^{-1} stereo radiography) has demonstrated skin marker motion relative to the femur averaging 1–5 mm throughout the motion, with peak-to-peak errors of the oscillation at impact averaging 7–14 mm (13). Errors were both subject and activity specific (14). Techniques have been developed for improving estimates of bone dynamics from skin markers using large numbers of markers and optimization–modeling of soft tissue deformation (15,16) but the performance of these methods for *in vivo* studies has only been validated for the tibia (during a slow, impact-free 10 cm step-up movement of a single patient with an external fixator). Average errors were low, but peak errors routinely exceeded 1 mm. Errors would likely increase significantly during faster movements and movements involving impact, and also where the soft tissue layer between skin and bone is thicker (e.g., for the femur). However, if the kinematic measurements are to be used in conjunction with musculoskeletal models to estimate dynamic loads and stresses on joint tissues, then even errors as small as 1 mm may be unacceptable. For example, when estimating strains in the ACL, a ±1 mm error in tibio–femoral displacement could introduce uncertainty in the ligament length of approximately ±3% (assuming a nominal ligament length of 30 mm). This error is similar in magnitude to estimated peak ligament elongation occurring during common activities, such as stair climbing (17). For investigating cartilage deformation, this error magnitude would be even less acceptable. A 1 mm displacement would be equivalent to a cartilage strain of $\sim 25\%$, relative to the average thickness of healthy tibio–femoral cartilage (18). A displacement error of this magnitude would translate into huge differences in estimates of contact forces. Thus, efforts to model, predict and correct for soft tissue deformation are unlikely to achieve sufficient accuracy for assessing soft tissue behavior. Alternatively, kinematics from a high speed stereoradiographic system capable of tracking implanted tantalum markers *in vivo* with 3D accuracy and precision better than 0.072 mm in translation and 0.35° in rotation (19) are more appropriate in use with advanced computational models. This accuracy is an order of magnitude or greater of improvement over conventional motion analysis techniques, and is uniquely capable of providing the accuracy necessary to model joint stresses.

From Experimental to Advanced Theoretical Analysis in Joint Mechanics

In addition to measuring joint kinematics and contact areas, investigators have attempted to measure articular contact stresses and pressures. However, stresses throughout the cartilage layer cannot be measured experimentally. Direct measurements of stress can be made at the articular surface using pressure sensing devices (20–22) (e.g., pressure sensitive Fuji film, piezoresistive contact pressure transducers, dye staining, silicone rubber casting). For cadaver studies, Fuji film sheets (Fuji Prescale Film; Itoh, New York, NY) are inserted in a joint and if pressed produce a stain whose intensity depends on the static applied pressure. Alternatively, digital electronic pressure sensors (e.g., K-scan, Tekscan, Boston, MA) can be placed onto the articular surface. These sensors are thin and flexible, and can be made to conform to the anatomy of the medial and lateral knee compartments. They consist of printed circuits divided into grids of load-sensing regions. Each load-sensing region within the grid has a piezoresistive pigment that can be used to determine the total compressive load within that region. After appropriate calibration procedures, dynamic pressure distributions can be calculated. In addition to providing a continuous, dynamic readout, K-scan has been reported to more accurately estimate contact areas than Fuji film (23,24).

There are significant concerns with the use of these sensors for estimating actual contact pressures. These techniques measure only surface-layer stresses, they alter the nature of cartilage surface interactions and are too invasive for *in vivo* human use. Thus, the clinical validity of articular pressure measurement with such sensors is questionable. They can, however, be important tools for the evaluation of the predictive power of joint models (2). By including the sensor in a finite element model, the effects of the sensor film on the actual contact mechanics can be accounted for (25). Thus, contact pressure predictions from such models can be directly compared to the pressure sensor measurements for finite element (FE) model validation.

Many *in situ* experimental studies have been conducted to obtain 3D knee joint kinematics and force-displacement data (21,26–30). Cadaver studies, however, cannot reproduce the complex loading seen by the joint during strenuous movements, since the muscle forces driving the movement cannot be simulated. Because of these fundamental limitations of experimental measures, mathematical models are favored for obtaining comprehensive descriptions of the spatial and temporal variations of cartilage stresses. A numerical model could be used to perform parametric studies of geometry, loading or material properties in controlled ways that would not be possible with tissue samples.

Theoretical Analysis of Joint Mechanics

During the last two decades, a number of theoretical joint mechanics studies with different degrees of accuracy and predictive power have been presented in the literature (31–39). Computational modeling work has included anatomical or geometrical observation (40,41) and analytical mathematical modeling (42–46). More recently, advanced FE modeling approaches allowed for improvements in the predictive power of localized tissue deformation (47–54). Joint biomechanics problems are characterized by moving contacts between two topologically complex soft tissue layers separated by a thin layer of non-Newtonian synovial fluid. A prime example is the multibody sliding contact problem between the tibia, femur, and menisci. The complexity of such problems requires implementation of sophisticated numerical methods for solutions (55–58). The finite element method is ideally suited for obtaining solutions to joint contact problems. Thus far, much of the finite element analysis has been applied to the study of hard tissue structures, often as it relates to prosthetic devices (59,60). When addressed, soft tissue layers are treated as single-phase elastic materials. As a consequence of the relative dearth of precise patient specific geometric data, material properties and insufficiency in accuracy of *in vivo* kinematics for input, no patient specific computational models have been reported for longitudinal clinical joint studies.

Surface Modeling

Surface modeling methods calculate the shape variations of joints and visualize the proximity of subchondral bone surfaces during static loading or dynamic movement. These methods can combine *in situ* data, motion analysis optical system data or high speed biplane radiographic image data and 3D bone surface information derived from computed tomography to determine subchondral bone motion. This method can be used to identify the regions of contact during static loading or dynamic motion, to calculate the surface area of subchondral bone within close contact, and to determine the changing position of the close contact area during dynamic activities (Fig. 9).

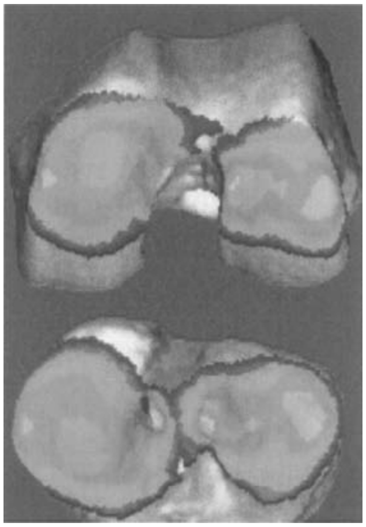


Figure 9. Example applications showing dynamic *in vivo* tibiofemoral bone surface motion using joint proximity (Euclidian distance) mapping during human one-legged hopping.

In vivo dynamic joint surface interaction information would be useful in the study of osteoarthritis changes in joint space and contact patterns over time, in biomechanical modeling to assist in finite element modeling, and in identifying normal and pathological joint mechanics pre- and postsurgery. Previous attempts to quantify the interaction between bones have utilized various methods including castings (62,63), pressure sensitive film (64), mathematical surface modeling (65,66), implant registration (67) and cine phase contrast magnetic resonance imaging (MRI) (68). The casting method can only be applied to cadaver models under static loading conditions. Pressure sensitive film also requires a cadaver model and necessitates inserting material into the joint space. Mathematical surface modeling allows analysis of dynamic motions *in vivo*, however, the joint must be disarticulated after testing. Implant registration requires either surgical implants or nonsubject specific image matching algorithms. Cine phase contrast MRI requires repeatedly performing the same motion pattern during testing and is limited to a small range of motion. The process described below is an improvement on these previous techniques because it utilizes live subjects performing dynamic tasks with unrestricted motion. Direct measurement of articular cartilage behavior *in vivo* during dynamic loading is problematic. In order to estimate the behavior of articular cartilage, the surface proximity interaction method that precisely tracks the motion of subchondral bone surfaces in vivo. Articular cartilage behavior is then estimated from these subchondral bone measurements.

Anderst et al. (61) described a method to estimate *in vivo* dynamic articular surface interaction by combining joint kinematics from high-speed biplane radiography with 3D bone shape information derived from computed tomography (CT). Markers implanted in the bones were visible in both the CT scans and the radiographic images, and were used to register the subchondral bone surfaces with the 3D bone motion. Joint surface interactions were then estimated by analyzing the relative proximity of the subchondral bone surfaces during the rendered movement. Computed Tomography data can be also used for joint geometry–shape characterization. The method is referred as reconstruction of volumetric models into rendered joint surface geometry models.

Computed Tomography data are typically collected for this method with slice spacing between the different images of $\sim 0.625 - 1.25$ mm and the in-plane resolution

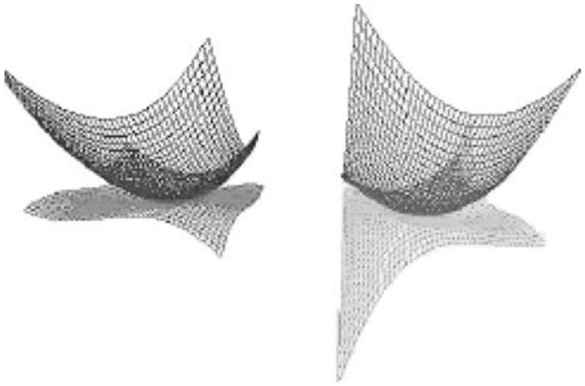


Figure 10. Articular surface matching (femoral condyle on top and tibial plateau below) using geometrical objects (69).

is ~ 0.293–0.6 mm depending on the size of the bone. The CT scans are reconstructed into 3D solid figures using software that employs reconstruction techniques, that is, the regularized marching tetrahedra algorithm by Treece et al. (70). If necessary, threshold values are adjusted to ensure the entire bone surface appeared in the reconstruction and the opposing bone surfaces never overlapped in computer animations of the motion.

Anterior–posterior and lateral radiographs are commonly used to preoperatively determine prosthetic size and proper donor selection for osteochondral allografts. By using 3D computer aided design tools and the reconstructed 3D joint geometry from CT described above, size determination is less prone to out-of-plane imaging errors associated with sagittal and coronal roentgenograms. Assessment of surface size, curvature analysis and knee incongruity is possible with *in vivo* CT [Fig. 10 (69,71,72)]. After the 3D joint surface reconstruction models the distal articular femur ($n = 16$) can be represented by six circles, the diameters of these circles, their angular arcs, and the distances between their centers varied with the size of the femur (Fig. 11; Table 5). There is a statistically significant association between several geometry parameters when the lateral or the medial distal femur is studied independently. These associations do not exist when we correlate medial versus lateral compartments across the population.

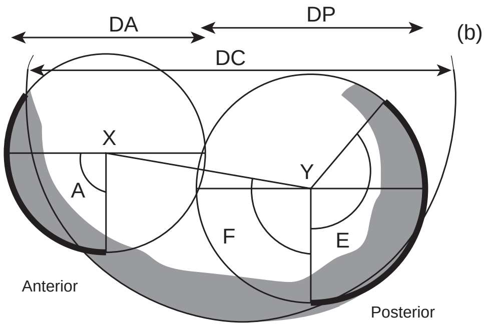


Figure 11. All the sagittal view measured parameters in the study of femoral head congruity (69).

The Joint Distribution Problem

Much attention has been devoted to the solution to what has become known as the general joint distribution problem that is, the problem of estimating the *in vivo* forces transmitted by the individual anatomical structures in the joint neighborhood during some activity of interest (73). The prediction of forces in joint structures has many applications. In the field of medicine, these predictions are useful for obtaining a better understanding of muscle and ligament function, mechanical environment within which prosthetic components must operate, and mechanical effects of musculoskeletal diseases. In the realm of sport biomechanics, these predictions are useful for better understanding of the kinetic demands, performance constraints, and mechanisms for improving athletic performance. Industrial applications include the optimization of occupational performance and safety considerations. Although the general techniques for predicting forces in joint structures may be used throughout this broad range of applications, the particular method of choice and the details of the analysis depend on the application.

The general force distribution problem normally arises in the following way. The musculoskeletal system or a relevant portion thereof is modeled as a mechanical system consisting of a number of essentially rigid elements (body segments) subjected to forces due to the presence of a gravitational field, and segmental contact with external objects, neighboring segments, and soft tissue structures that produce and constrain system motion. The associated inverse dynamics problem is then formulated and solved to determine the variable intersegmental (joint) force and moment resultants during the activity of interest. The joint resultants are abstract kinetic quantities that represent the net effect of all the forces transmitted by the anatomic structures crossing the joint. At a typical joint, these forces normally include the forces transmitted by the muscles, ligaments, and articular (bony) contact surfaces.

The unknown forces transmitted by the joint structure are next related to the known intersegmental resultants by writing the joint equilibrium equations. These equations express the fact that the vector sum of all the forces in the individual anatomic structures, and the vector sum of all the moments about the joint center produced by those forces, are equal to the intersegmental resultant force and moment, respectively. Assuming that all joint geometry (point of application and orientation of forces) is known and that these two independent vector equations (or six independent scalar equations) involve as unknowns the M muscle and L ligament forces, together with the $3C$ scalar components of the C bony contact forces, these joint equilibrium equations are indeterminate whenever the sum ($M + L + 3C$) of the unknown forces exceeds six. Thus, if the system model includes only one bony contact force ($C = 1$) and more than three muscle and/or ligament forces ($M + L > 3$), the corresponding joint distribution problem will be indeterminate and therefore have an infinite number of solutions.

Finally, the joint resultants are decomposed or distributed to the individual joint structures at each instant of

Table 5. Femoral ($n = 16$) Medial and Lateral Compartment Measurements[a]

	Medial	SD	Lateral	SD	Ratio M/L
DC mm	68.28	5.003	67.839	5.865	1.006
DA mm	42.45	10.086	44.41	4.608	0.955
DP mm	40.364	1.231	41.212	3.069	0.979
XY mm	24.519	2.686	23.529	3.069	1.042
F°	100.374	10.572	102.631	5.834	0.978
E°	151.168	10.9	139.629	12.509	1.0824

interest during the activity, using some appropriate solution methodology.

The general joint distribution problem may thus be stated in the following way. At any instant of time when the joint resultants are known, the forces transmitted by the individual joint structures are determined such that the equilibrium equations, and all relevant constraints on the forces in the individual joint structures, are simultaneously satisfied. The classical studies of joint distribution problems use essentially two different methods to solve the indeterminate joint distribution problem: the reduction method, and the optimization method.

The mathematical modeling of human anatomy and its functions has been influenced by two main simulation approaches or philosophies. In the first the joint structures are of no importance in the mathematical modeling while in the second simulation of the geometry and structural relationships of the joint components in addition to their behavioral properties are the main tasks. Hefzy et al. categorized these different approaches as phenomenological and anatomical, respectively (74).

Phenomenological Joint Models. The phenomenological models include two groups: the rheological models and the advanced figure animation models. The rheological models analyze the dynamic behavior of a system by treating it as viscoelastic, being composed of springs and dampers. However, the noncorrespondence of these components to the structure of the components in the knee leads to no structural information in the model output.

The advanced figure animation models provide information on body dynamics by taking into account body segment dimensions, masses, moments of inertia, and so on, but do not model the detailed geometry of joints.

Anatomical Models. Two different approaches to modeling categorization exist in the experimental literature. According to the first, the categorization of the models depends on the type of motion reproduced by the mathematics. The second approach categorizes models according to their structural basis. There is a vast number of studies attempting to model specific component structure and behavior that will be evaluated in order to identify the optimum method for the modeling of each specific component.

The Reduction Method. The reduction method reduces the number of unknown forces to correspond with the number of equations governing the distribution problem (or increases the number of equations to agree with the number of unknowns, e.g., the deformation-force relations for the unknown forces). For the general 3D distribution problem, this implies that the number of unknown scalar forces ($M + L + 3C$) must be reduced to six to allow for a unique solution. Previous investigators have reduced the number of unknowns by (*1*) grouping muscles and ligaments with apparently similar functional roles, (*2*) grouping multiply connected bony contact force regions, (*3*) assuming a direction for the unknown bony contact force, (*4*) using EMG data to determine when a muscle is active, (*5*) ignoring ligament forces except near the limits of the range of motion, and (*6*) ignoring antagonistic muscular activity. Several models [(73,75); Fig. 12a, b] predicted muscle forces and the bony contact force at the hip during locomotion. In these studies, the indeterminate distribution problem at the hip or the knee was made determinate through several simplifying assumptions. The hip muscles were combined into six functional groups (long flexors, short flexors, long extensors, short extensors, abductors, and adductors), and ligament function at the hip was assumed to be negligible. The forces transmitted by these six muscle groups, combined with the three components of bony contact force, comprised nine unknown scalar quantities. Only two of the three components of the resultant hip moment were considered; analysis of the component tending to internally or externally rotate the femur relative to the pelvis was rejected as inaccurate. Previous reports of EMG activity were to demonstrate that there is little antagonistic muscle action, and only muscle agonists were considered. Despite these simplifications, the possibility of activity in both the long and short flexors and extensors still made the problem indeterminate. A solution was obtained, however, by assuming activity in only one flexor (either long or short, but not both) and one extensor.

The Optimization Method. The previous discussion indicates that the distribution problem at a joint is typically an indeterminate problem, since the number of muscles, ligaments and bony contact regions available to transmit forces across a joint in many cases exceeds the minimum number required to generate a determinate solution to the joint equilibrium equations. Determinate solutions are obtainable only with significant simplification of joint function or anatomy. In contrast, the optimization method of solving the general joint distribution problem does not require such simplification. Rather, it retains many of the anatomical complexities incorporated in defining the problem, and seeks an optimum solution (i.e., a solution that maximizes some process or action). Optimization techniques may be divided into linear and

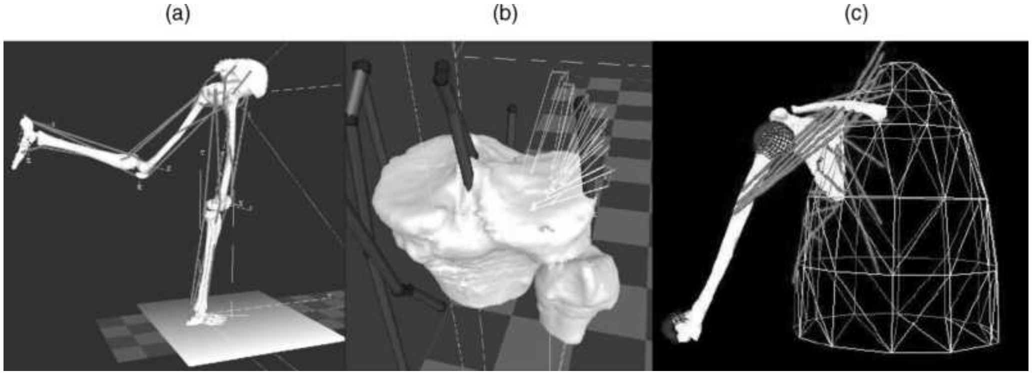


Figure 12. Forward and inverse analysis driven mathematical models, (a) the Strathclyde model animated using SIMM Musculographics-Motion Analysis Software. (From Ref. (77)) (b) Force distribution method-inverse dynamics, the vector of bony contact force is shown on the tibial plateau. (From Refs. (71) and (77)). (c) the Delft forward and Inverse shoulder model (78).

nonlinear methods Simultaneous use of linear cost and constraint functions in the model formulation constitutes the so-called linear optimization method. In contrast, if the cost function and/or one or more of the constraint functions is nonlinear, solution of the problem requires the use of nonlinear optimization methods. Both of these optimization methods have been made practical with high speed computers, since the techniques require many iterations of equation-solving to find an optimum solution. Neuromuscular function is complex and poorly understood at the conceptual, if not detailed level. However, it is generally assumed that physiological functions are optimized in some way. In 1836, the Weber brothers commented that we walk and run in "the way that affords us the least energy expenditure for the longest time and with the best results". Experimental evidence suggests that oxygen consumption (and presumably energy expenditure) is minimal at freely selected walking speeds, supporting the assumptions of optimal physiological neuromuscular function.

The optimization criteria that the neuromuscular control system "chooses", either consciously or unconsciously, to select muscle action may vary considerably with the nature of the physical activity to be performed and the physical capabilities of the individual. For example, muscle control in sprint running may serve to maximize velocity, while in walking the control process may serve to maximize endurance. In a painful pathological situation, such as degenerative joint disease, muscular control may serve to minimize pain. If this pain occurs due to the joint surface pressure, the appropriate optimization criterion may be to minimize to bony contact force. Muscular control may also serve to minimize the forces transmitted by passive joint structures such as ligaments. These examples indicate that there are possible optimization criteria to choose from, and the choice of a criterion to solve a particular distribution problem may not be obvious.

Given that gait presents a relatively unambiguous performance criterion, one can claim that it fits into the framework of optimum control theory. Additionally, gait presents a characteristic bilateral symmetry that leads to relatively simple representation of the dynamic system. Activity in the stance phase of gait is described by equality constraints on the "states", knowing that each gait stage involves dynamic constraints that reflect the particular nature of the phasic activity. However, our motivation in the use of optimum control for the study of movement relies upon the belief that it is currently the most sophisticated methodology available for solving extremely complex problems. Optimal control theory requires not only that the system dynamics are precisely determined and formulated but also that an appropriate performance criterion is chosen. Therefore, deficiencies in the modeling that are present in either the system dynamics or the performance criterion are indicated by differences between model and experiment.

Forward Analysis. The technique of forward simulation allows the study of the causal relationship between forces acting on a mechanical structure and the resulting movement (79). A first approach requires the description of joint moments, initial positions and velocities for each body segment as input variables. The forward dynamics problem then is expressed as a set of differential equations of motion with associated restraints and solved to yield angular accelerations. Angular velocities and displacements are then determined by integration (80). Modifications to input variables, either joint moment profiles or initial segmental configuration, can then be introduced and the resulting changes in the movement pattern evaluated.

Due to recent improvements in musculoskeletal modeling, forward simulation driven by muscle activity rather than joint moments is now possible. The definition of muscle activation sequences and the initial configuration of the mechanical system (angles and angular velocities) are entered into the forward simulation. A physiological model describing muscle excitation characteristics as well as muscle mechanics is used for calculating the individual muscle force production and the resulting motion is calculated.

Because of the inherent complexity of the musculoskeletal model and the multiple interaction of parameters, forward simulation alone is unlikely to contribute to the identification of aberrant muscle action which affects timing or force production, causing an aberrant gait pattern. Matching the system response to the joint movement profiles observed in an individual patient would require extensive trial and error with no guarantee of ever finding the parameter setting responsible for the aberrant joint movement.

Optimization has been used previously as a curve-fitting tool in addition to forward simulation techniques to reproduce the movement pattern observed in an individual subject. In the studies of Chao and Rim (81), and also Townsend (82), optimization techniques were used to determine the applied joint moments by iteratively varying them until the theoretical limb displacement fits those measured in the laboratory. Chou et al. (83) predicted the minimum energy consumption trajectory of the swing limb using a similar approach. Using approximate muscle-force and/or joint moment trajectories, Pandy and Berme (84,85) evaluated body segment motion and ground reaction forces during single stance, based on a 3D model incorporating 7 DOF. In Jonkers et al. (76), initial inputs to the forward simulation process were the normalized quantified muscle activation patterns of 22 muscles, and the initial segmental configuration (both angles and angular velocity) derived from Winter (86). Two distinct musculoskeletal models (one including 6 DOF, the other 7 DOF) were defined and a muscle driven forward simulation was implemented. A series of optimization sequences then were executed to modify the muscle activation patterns and initial segmental configuration, until the system output of the forward simulation approximated the angle data reported by Winter (86). The accuracy and effectiveness of the analysis sequence proposed and the model response obtained using two distinct musculoskeletal models were verified and analyzed with respect to the kinesiology of normal walking.

Based on the integrated use of optimization and forward simulation techniques, a causal relation between muscle action, initial segmental configuration and resulting joint kinematics during single limb stance phase of gait was successfully established for a musculoskeletal model incorporating 22 muscles and 7 DOF (Fig. 12). Despite the inherent simplifications of the planar models used in this study, several kinesiological principles of normal walking were confirmed by the analysis and a reference base for exploring the causal relation between muscle function and resulting movement pattern was established.

Finite Element Analysis of Human Joints. Adaptation of finite element analysis (FEA) to the analysis of stress in human joints, requires fundamental studies in the following areas: (*1*) three dimensional FEA of moving contact problems utilizing finite deformation laws (linear elastic versus biphasic) for cartilage and non-Newtonian laws for synovial fluid; (*2*) automated adaptive methods for the generation and control of 3D computational models using error estimates and controls that account for nonlinearities, singularities, and boundary layer effects; (*3*) improvements in geometry of FE models by using patient specific MRI and CT imaging data. Recently available semiautomatic 3D mesh generation tools (reconstructed mesh from the volumetric imaging data), and numerical methods for processing the material and geometric data required for the contact analysis considerably reduce the time requirements of such efforts; (*4*) implementation of recently available high accuracy 3D joint *in vivo* kinematics to calibrate the FE models, and to assist in simulating strenuous activities; (*5*) Parallel solution algorithms for the nonlinear time-dependent problems utilizing high performance computer architectures.

Several models of articular cartilage have been proposed to describe its mechanical behavior; however, none of these models have been able to address the full spectrum of this tissue's complex mechanical responses. Typically, these models implement a subset of known tissue behavior, for which material properties must be determined from experiments. Experimental results indicate that cartilage exhibits flow-dependent viscoelasticity, anisotropy, and tension compression nonlinearity due to its ultrastructure and composition (87). These characteristics are further compounded by the inhomogeneity of the tissue through its thickness. It is widely accepted that the time-dependent response of cartilage can be accurately represented by the biphasic theory derived by Mow et al. (21). Under isotonic conditions, this biphasic theory of incompressible solid and fluid phases is appropriate for most applications involving cartilage modeling for infinitesimal or finite deformations (26,88). Numerical methods are required to solve these nonlinear problems, even for relatively simple geometries. Linear 3D elements (89) and nonlinear 3D formulations have been presented for the biphasic theory (90–92). However, full nonlinear 3D analysis for joints of realistic geometry and strains remains a computationally challenging problem.

Hirsch (1944) proposed to use the Hertz contact theory for contacting elastic spheres to model cartilage indentation (28). Askew and Mow (1978) analyzed the problem of a stationary parabolically-distributed normal surface traction acting on a layered transversely isotropic elastic medium to assess the function of the stiff surface layer of cartilage (29). More recently, complex, biphasic models have been developed that target slow, quasistatic loading response (89,90,93–95). Under high loading rates, however, the cartilage demonstrates a mechanical behavior that does not deviate substantially from a linear elastic model (38,96), so complex and numerically unwieldy biphasic models may not be necessary for rapid loading events. Eberhardt et al. (1990) (36) and (1991) (97) developed a solution for the contact problem of normal and tangential loading of elastic spheres, with either one or two isotropic elastic layers to model cartilage. A modified Hertzian (MH) theory that takes into account the thickness effect of the elastic layer has been used for articular joint contact analysis by several investigators (23,32,37,97). The results from the studies listed above have demonstrated that considerable differences may be found in cartilage stress predictions depending on the particular cartilage constitutive model being employed, that is linear elasticity theory versus linear biphasic theory. In view of the above, it has been proposed that it is adequate to use a transversely

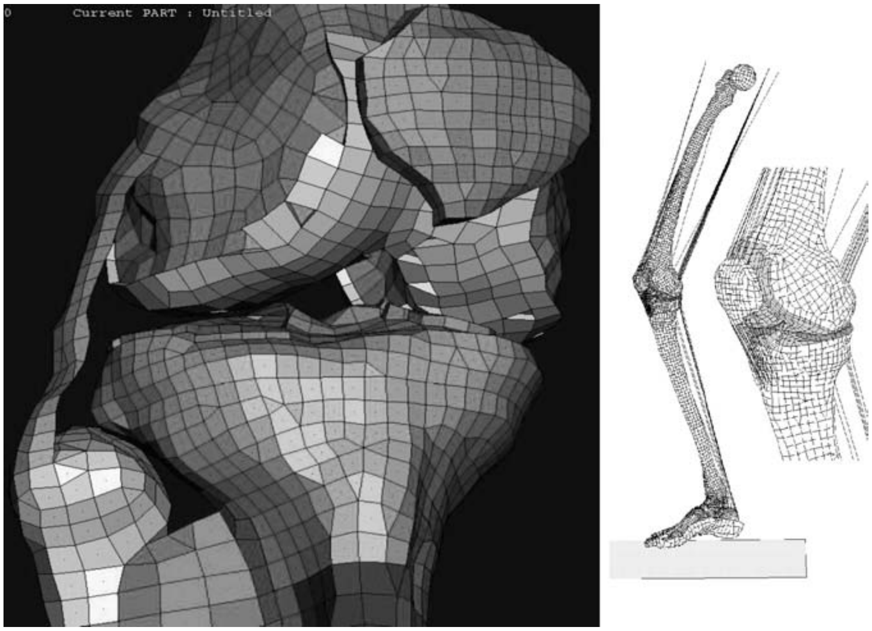


Figure 13. (a) Knee FE model components shown in different color-implicit formulation (102) (b) the knee FE model during simulation of single leg-hopping-explicit formulation (103).

isotropic, linearly elastic, homogeneous constitutive relationship to model the normal contact pressure distribution of the tibial plateau (33,98).

In the case of FE modeling applied to impacts and very rapid movements, model formulation problems must be accounted for by the solution. Implicit techniques are not widely used as they can be limited by problems of convergence. This method does not use kinematics to verify whether contact occurs at the calculated points at the associated load levels. For those applications, explicit techniques have proven useful in their ability to simulate deformations and contact conditions simultaneously as well as large segmental displacements (70,99–101) (Figs. 13b and 14). Explicit finite element analysis was also proven to be a valuable tool when simulating total knee replacement motions due to loads applied by a knee simulator (44,45). Stability of solution and low computational cost are the two main advantages of the explicit methods over classic implicit techniques when forces are used to drive the models (44,98,103–109). The main limitation of the explicit technique is that the method is only conditionally stable. Combination of both techniques is suggested so that the explicit formulation will be used to "educate" the implicit algorithms. Implicit formulations are used when the localized effect of articular stress needs to be addressed at the cartilage level with increased subject specific refined geometry since these techniques are more appropriate for the task (Fig. 13a).

Toward Patient-Specific and Task-Dependent Morphological FE Models. In order to apply FE computational methods to problems of joint mechanics, precisely measured anatomical data must be used to construct a 3D solid model from which the appropriate finite mesh can be constructed and analysis performed (110,111). Computed tomography (112) and MRI (113,114), or reshaped digital calipers and machine–controlled contact digitization (23) are commonly used methods to obtain soft–hard tissue geometry. Clinical modalities of these imaging techniques have resolutions that are typically limited to 500 μm. Considering that the articular cartilage of the knee joint is only ~ 4 mm thick (14), a resolution of 500 μm is generally inadequate and special protocols are required for the imaging procedure (19). Note that all properly calibrated 3D FE knee joint models to date, have been developed from images of cadaveric knees (43,115). Models with patient specific geometry are possible using

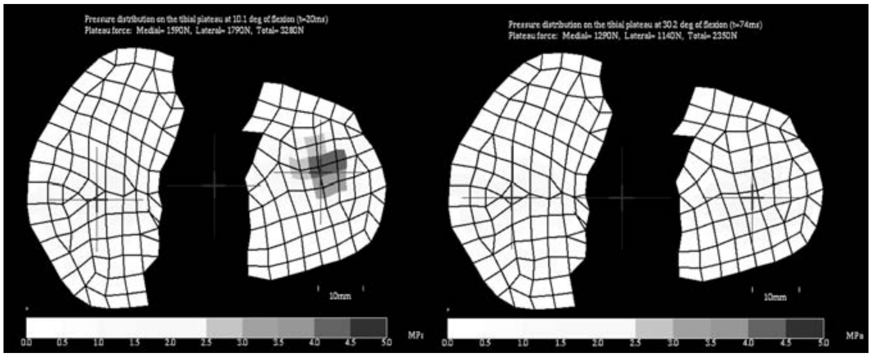


Figure 14. Tibial plateau pressure at 10 and 30 degrees flexion during single-legged hopping (103).

combined appropriate MR and CT imaging modalities that will result in increased model detail (100).

With recent technological advances in 3D imaging, computer hardware, and FE formulations, it has become possible to estimate human tissue properties (100,103). Estimation of the stiffness of hard tissue (i.e., cancellous bone) using large-scale finite element (LSFE) models or microFEs at both the apparent and trabecular levels is therefore possible (47,52,70,116–122). This direct analysis approach has advantages over traditional statistical methods in that (*1*) all morphometric measurements available from the 3D image data are contained in the FE model, and (*2*) the FE model of the structure is a mechanistic rather than statistical approach to the prediction of biological tissue properties. Application to these techniques in the clinical environment with *in vivo* data is gaining considerable ground. These recent improvements in imaging, allow the generation of 3D models in a semi-automatic fashion (23,69,123,124). Models now provide the opportunity to study the mechanical effects of individual geometrical variation (125). Such ***subject specific finite element models*** are mainly useful under two conditions: (*1*) In the statistical analysis, if the differences between subjects are large relative to the differences in model predictions related to the choice of input parameters that can only be estimated. (*2*) In a contact FE problem with high accuracy 3D kinematics refined meshes are needed in the vicinity of the contact zone (114).

The inherent ability of FEM to overcome geometrical and topological modeling problems is no longer burdened by some of the limitations apparent in early studies. Such limitations have been overcome by recent computational and experimental advances. These limitations included insufficient accuracy of *in vivo* kinematics–kinetics input data, inadequate hard–soft tissue imaging, excessive computational cost and the inherently complex process of material characterization in relation to some of the components of the stress tensor in biological material.

JOINT STABILITY

Newton's third law states that forces always exist in pairs that are equal and opposite in direction, such that if one body pushes against another, the second body will push back against the first with a force of equal magnitude if the state of motion is constant. This would mean that in order for muscles to be able to pass movement on to a limb segment, there has to be an interaction between the segment and another bone, and a joint structure that will allow the desired rotational direction and force. Such a mechanism prevents translational movements, which can "dislocate" one rigid bony limb segment from another.

Bones, ligaments, and muscles all contribute to joint stability. The extent each structure participates in maintaining joint stability differs between joints. Because bones are rigid relative to the other tissues in the joint, they can provide great stability to the joint. In general, the greater the circumference of the bony segment enclosed by its counterpart, the greater the amount of translational stability that exists in the joint. For example, the femoral head is almost completely enclosed by the acetabulum in the hip joint, whereas the humeral head is only slightly enclosed by the glenoid in the shoulder. It is obviously easier to dislocate the shoulder than the hip.

Ligaments, much like cables, can restrict rotational or translational motions depending on their location and the direction of the force. For example, the cruciate ligaments limit the anterioposterior translation of the tibia on the femur at the knee joint, while talocalcaneofibular ligaments prevent rotational motions as well as translational motions between talus, fibula and calcaneus. However, ligaments, unlike bones, cannot provide rigid constraints as they are relatively soft and flexible.

Muscles–tendons are also soft and flexible and, like ligaments, provide nonrigid constraints to the joint. A significant difference between the two, however, is that muscles are active in controlling the joint motion whereas ligaments are only passive stabilizers. Muscle action is usually complementary to that of ligaments and, in fact, can protect ligaments from damage or can protect the joint from further damage in case that the ligament is ruptured. Tensile forces exerted on muscles are counterbalanced by compressive forces across the joint surfaces providing stability to the joint against forces acting to open up the joint space. Thus, muscles can support or limit motions and serve the dual function of providing desired movements while contributing to joint stability.

The Hip Joint

The hip joint is the link of the upper body and the pelvis–trunk with the lower limbs, the main locomotion facility of the body. It is a ball-and-socket joint (Table 4) in which the head of the femur resides in the acetabulum of the pelvis, making one of the largest and most stable joints in the body. The surface area and the radius of curvature of the articular surface of the acetabulum closely match that of the articular surface of the femoral head. The hip joint is structurally a highly constrained joint. Because of the inherent stability conferred by its bony architecture, this joint is well suited for performing the weightbearing supportive tasks that are imposed on it.

The femoral ball is embraced by the acetabular socket, allowing rotation to occur with virtually no translation. The cartilage that covers the acetabulum thickens peripherally (126). A plane through the circumference of the acetabulum immediately at its opening would project with the sagittal plane intersections at an angle of 40° (opening posterior) and 60° (opening laterally). This architectural constraint imparted by the bony shapes almost eliminates the need for ligamentous and soft-tissue constraints to maintain the stability of the hip articulation. Although this increased constraint provides stability to the hip, there is a structural drawback. Such constraint limits the global range of motion of this joint at the fulcrum of the lower extremity. Fortunately, human biomechanics and everyday tasks performed by the hip do not violate these limitations and the hip's range of motion is very rarely subjected to extremes. During most ambulatory activities, such as normal bipedal locomotion, the lower extremity is positioned anteriorly in the sagittal plane with only small rotations necessary in the other two planes. Hip flexion of at least 120°, abduction of at least 20°, and external rotation of at least 20° are necessary for carrying out normal daily activities. Activities such as descending stairs sitting, rising from a chair, and dressing require greater degrees of flexion and rotation at the hip joint. For example descending stairs requires 36° of motion whereas squatting requires 122° of motion in the sagittal plane (127).

The hip joint reaction force (at a neutral position) can reach three times body weight (BW) in single legged stance, but could get up to six times BW during the stance phase of gait and increases significantly with gait velocity. The main mechanism influencing this magnitude is the ratio of the abductor muscle force and the effect of the gravitational force moment arms. The rule normally suggests greater reaction forces expected at low ratios (128). Bracing and use of a cane can decrease the hip joint force.

The Knee Joint

The knee consists of a two joint structure: the femorotibial joint and the patellofemoral joint. The femorotibial joint is the largest joint in the body and is considered to be a modified hinged joint containing the articulating ends of the femur and tibia. The patellofemoral joint consists of the patella, the largest sesamoid bone, and the trochlea of the femur. Taken together, the knee joints function to control the distance between the pelvis and the foot as a control link. Because of the role of the knee in weight-bearing, its surrounding of very strong musculature and its location between the two longest bones in the body, tremendous forces are generated across it. Surface motion occurs simultaneously in both sagittal and the transverse plane with the first being the dominant plane of motion (129). During activities such as running, landing, and pivoting, the knee functions to maintain a given leg length and acts as a shock absorber. During stair climbing, crouching, and jumping, large propulsive forces at the knee (several times BW) are generated to control the degree and speed of shortening and lengthening of the leg. In these situations, knee stability is a dynamic process maintained through fixed bony and ligamentous constraints and modified by the action of the muscles crossing the joint. The higher the rate of the loading the more demanding the role of the knee in sustaining stability. The bony morphology of the femur, tibia, menisci, cruciate–collateral ligaments, and patella contribute to joint stability, but to a lesser extent than that of other more constrained joints such as the hip. Static constraints play a very significant role in knee stability as compared with the shoulder for example where stability is maintained through the dynamic action of the surrounding muscles. The cruciate and collateral ligaments are the major structures limiting motion at the knee. The posteromedial and posterolateral capsular complexes augment the four primary ligaments, with the menisci playing a lesser role. The muscles crossing the knee contribute to dynamic stability and are particularly important in the presence of pathologic laxity. A method that describes the constraining mechanism of the anterior and posterior ligament has used the notion of the 2D "four bar linkage". The instant center of rotation, designated primarily by the femoral condyle surface shape, follows a semicircular pathway, and the direction of displacement of the femorotibial contact points is tangential to the surface of the tibia, indicating gliding throughout the range of motion. However, the axis of rotation at the knee does not remain fixed during flexion. Indeed, as the knee flexes the screw axis will sweep out a ruled surface in space, known as the axode. This fluctuation in the screw axis signifies that the knee is not truly a hinge joint, for which the axode would degenerate to a fixed line in space. Usually, the knee is approximated as a hinge joint, a simplification that may be acceptable for flexion angles between 45 and 90° where the moving screw axis remains very close to the line passing through the centers of curvature of the two posterior femoral condyles. The motion at the articular surfaces is not one of pure rolling, but a combination of rolling and sliding as indicated by the screw axis that never lies near the articular surfaces of the tibiofemoral joint.

The Foot Structure; the Ankle Joint

The ankle joint can be described as a saddle-shaped lower end structure of a long bone (tibia and fibula) Fig. 15. Its inferior transverse ligament encloses the superior aspect of the body of the talus (the trochlea). It is the joint that first receives the transient impact that travels through the tibia in gait or other movement. It alternates in both form and function to receive load as a shock absorbing

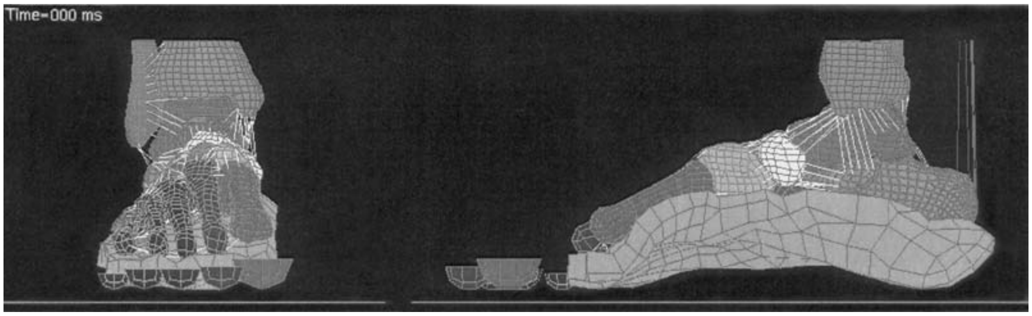


Figure 15. The Wayne State University Ankle joint and foot Finite element model (130).

mechanism and to propel significant leverage based force during fast locomotion. The subtalar and ankle joint act like a mitered hinge. The tibial surface forming the superior dome of the ankle is concave sagittally, is slightly convex from side to side, and is oriented ~ 93° from the long axis of the tibia (it is higher on the lateral than the medial side). The upper articulating surfaces of the talus appear to match closely that of the cavity formed by the tibiofibular mortise. The superior part of the body of the talus is wedge-shaped. It is about one-fourth wider in front than behind, with an average difference of 2.4 mm; anteriorly a minimal difference of 1.3 mm and a maximal difference of 6 mm. From front to back, the articular surface spans an arc of ~105°. This surface contour, having a smaller diameter medially than laterally, has been compared to a section or rostrum of a cone. The primary motion of the ankle joint is dorsiflexion–plantarflexion. Its axis of rotation is obliquely oriented with respect to all three anatomic planes with ankle dorsiflexion and tibial internal rotation being associated with subtalar eversion (pronation) whereas the ankle plantarflexion and tibial external rotation are associated with subtalar inversion (supination). The axis extends from anterior, superior, and medial to inferior, posterior, and lateral as it is passing through the inferior tips of the malleoli. It is at angles of 93° with respect to the long axes of the tibia and ~12° to the joint surface. However, rather than a true single ICR, the ankle has been noted to have multiple instant centers, all of which perturbate and fall very close to a single point within the body of the talus. Through a complete arc of ankle rotation, the center may be displaced anywhere from 4 to 7 mm. The oblique orientation of the axis of rotation to the sagittal, coronal, and transverse planes, translation of the talus in the mortise can occur in all three directions. The talus has been observed *in vitro* to rotate easily in the ankle mortise implying relative movement between the malleoli. Because the trochlea is wider anteriorly than posteriorly, it has been suggested that lateral play of the talus within its mortise occurs only when the ankle is in plantarflexion. Subtalar motion has been described as screw-like influencing the flexibility of the transverse tarsal joint. Others suggest that instability exists in dorsiflexion, while others yet believe that with intact ligaments translation, occurs only in the sagittal direction. These differences can be explained by behavior of >100 ligaments and by the roles played by the subtalar joint, the kinematic chain of the hindfoot, and the muscles that traverse this area in transmitting forces across this area during plantarflexion and dorsiflexion. The talus is unique because this bone has seven articulations that connect it to four other bones, it lies between the foot and the leg, and contains no muscular attachments. The stability of the talus and its articulations, therefore, relies heavily on the ligamentous attachments and musculotendinous complexes that traverse the talus and attach distally. Therefore the main characteristic of ankle joint is its strong passive stability attributed to a variety of factors. First is the bony stability provided by contact of the trochlea with the tibial plafond. Second are the medial and lateral cartilaginous slightly concave surfaces that articulate with the two malleoli. Third are the ligamentous connections between the tibia, fibula, talus, and calcaneus. Ankle stability increases during weight bearing and depends more on articular surface congruency.

The tarsometatarsal joints are relatively mobile and intrinsically stable joints that produce the arch-like configuration allowing wide range of motion at the first metatarsophalangeal joint with gliding during most of its range and jamming artful extension. The medial longitudinal arch functions as a beam and a truss.

The mode of foot–ankle mobility and muscular control is the most significant determinant of both limb stability and body progression (131).

Standing barefooted we load our heels with two-thirds of the load while the other third is loading the forefoot. During walking and at the early phase of stance, the center of pressure moves from the posterolateral heel rapidly across the midfoot, a phenomenon coupled with the firing of the anterior tibial musculature to slow foot plantarflexion and prevent foot slap. Then, at mid- and late stance the posterior calf musculature fires, propelling the body over the foot towards toe-off phase where the hallux bears the most pressure.

At higher speeds/rates of mobility it is the intrinsic mechanical properties that provide control rather than the neuromuscular control system. The force at the ankle joint can reach magnitudes up to six times body weight during walking and thirteen times BW during running. The heel fat pad is a very effective shock absorbing mechanism and it has been shown that high heels, or narrow shoes, narrow toebox, can lead to altered foot mechanics that ultimately result in foot deformities and pain.

The Spine

The spine is composed of a series of vertebrae (7 cervical, 12 thoracic, 5 lumbar, the sacrum and coccyx) and intervening soft tissues, such as intervertebral disks, ligaments, and muscle attachments. It provides important functions including support of the body structure and protection of vital tissues such as the spinal cord, nerves and arteries. Yet it is flexible and allows mobility to the torso. Several studies have described the passive and active range of motion of spinal segments differently (Table 4). The motion of the spine is usually analyzed through consideration of motion segments that are comprised of two adjacent vertebrae with the intervertebral disk and other intervening soft tissues. These structures, also called functional spinal units (FSU), move with 6 DOF, however, the motion is quite complex due to six articulate faces between the two bony segments and attachment of multiple ligaments and muscles. Normally simultaneous translations and rotations of FSU are coupled in the analyses (134). Creep, relaxation and hysteresis are the three prevailing viscoelastic characteristics of the intervertebral disks (1). At high rates of loading the disks serve as a shock absorber with compression strength being higher from upper cervical to lower lumbar levels. Although this is true for most joints, vibrational properties of spinal segments have attracted particular attention as they are thought to be related to injury and pain. The spinal cord exhibits some longitudinal

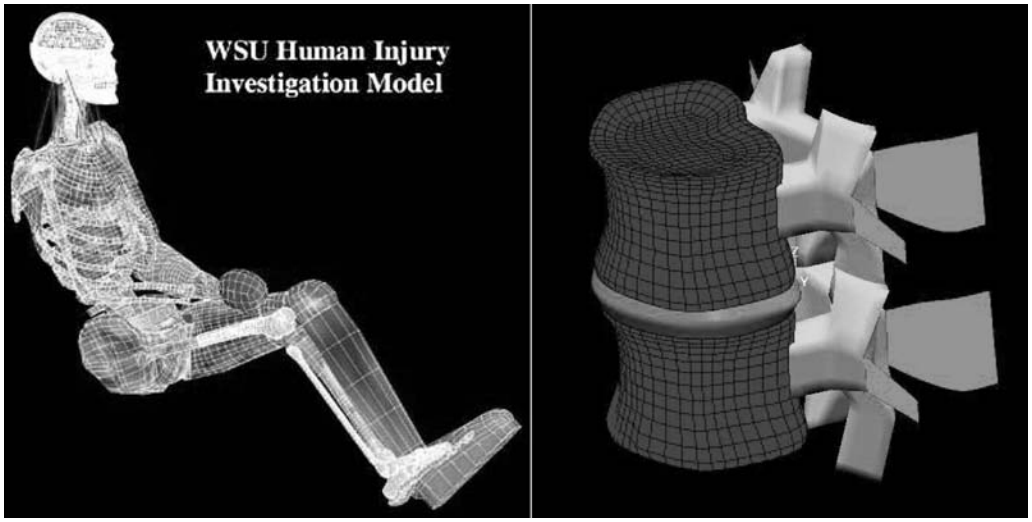


Figure 16. The Wayne State University full body impact model and the spine model (130,135).

elasticity but very poor axial translation. These translational forces are the ones primarily associated with neurological injury. Apparently muscular support is of vital importance is stability. If for example bilateral facet resection exceeds 50% the significant increase in annulus stresses may occur. Seatbelts can adequately protect front seat occupants against frontal impact in an airbag equipped vehicle (135).

Several models of the spine have been developed (130,135). A nonlinear finite element model of lumbar spine segment L3–L5 is shown in Fig. 16. The effects of upper-body mass, nucleus injury, damping, and different vibration frequency loads were analyzed for the whole body vibration (WBV) using this model. Anterior regions of L3–L5 segment show small vibration amplitudes, but posterior regions show large amplitudes. The posterior regions of intervertebral discs of lumbar spine are easy to injury during long-term WBV compared with anterior regions. The vibration of the human spine is more dangerous to facets, especially during WBV approximating a sympathetic vibration, which may lead to abnormal remodeling and disorders of lumbar spine.

The Shoulder

The shoulder complex allows the arm to move with respect to the thorax. The biomechanics of the shoulder involves the study of four different articulations, namely, the acromioclavicular, sternoclavicular, glenohumeral joints, and the scapulothoracic joints. The shoulder complex has the greatest range of motion among joints in the body (Table 4). Traditional descriptions of humeral motion are based on the angle between the humerus and the thorax in the sagittal plane (flexion and extension) and the coronal plane (abduction), with one very common characteristic: the glenohumeral articulation is inherently unstable due to the shallow nature of the glenoid fossa, containing only one-third of the diameter of the humeral head. The ligaments, capsular and muscular structures are the main stability mechanisms. Axial rotation of the humerus is conventionally described by the degrees of internal or external rotation when the humeral axis is parallel to the thorax, or perpendicular to the thorax (abducted 90°). The primary anterior stabilizers of the shoulder when the arm is abducted at 90° are the inferior glenohumeral ligaments. Horizontal abduction and adduction, also known as horizontal extension and flexion, respectively, are commonly used to describe arm position when the axis of the humerus is perpendicular to the thorax. The muscles surrounding the structure produce a barrier effect by applying compressive forces and by eccentric contraction resulting is active stability. The level of elevation of the normal arm is 167–168° and 171–175° for men and women, respectively. Average extension or posterior elevation is approximately 60°. When the arm is adducted by the side of the body, it is possible to achieve approximately 180° of rotation. With abduction of the arm to 90°, however, the total arc of rotation is reduced to 120°. The range of motion of the shoulder decreases with normal aging. The maximum loads that are characteristic for the glenohumeral joint can reach magnitudes of one and a half body weight (132).

The Elbow

The elbow joint consists of the proximal end of ulna and the capitulum of the humerus. Three articulations are present: Humeroulnar, humeroradial, and proximal radioulnar (133). It is a hinge-type joint whose functions include positioning and stabilizing the hand, providing lever-type support during lifting and weight-bearing during activities, such as a pommel horse exercise. The humerus-ulna joint itself has only one degree of freedom, however, the rotational motion of the radius over ulna provides another degree of freedom to the elbow. Thus the elbow can be considered to have two degrees of freedom. Basically its changing center of rotation during flexion-extension makes it more than a hinge joint: flexion-extension (0–160°)

and axial rotation, or pronation–supination (70–80° and 80–85°, respectively). The functional range of motion for most activities of daily living is 100° for flexion–extension (30–130°) and pronation–supination (±50°) at the elbow joint. The instantaneous axes of rotation for flexion–extension are localized at the center of the lateral projected curvature of the trochlea and capitulum. The size of this region measures only 2–3 mm.

The axis of forearm rotation passes through the capitulum and head of the radius, extending to the distal ulna. The carrying angle, the angle between the long axis of the humerus and the long axis of the ulna, averages 7° in men and 13° in women.

Motion in the plane of varus–valgus rotation at the elbow indicates joint instability and such motions are restricted. The rotational forces are mainly resisted by the anterior and posterior oblique fibers of the medial collateral ligament; the former is stretched during flexion and extension, whereas the latter is stretched only during flexion. In addition to the medial collateral ligament, the shape of the articular surface and the anterior capsule contribute to the resistance to valgus stress while the articular congruity and the radial head provide stability in flexion. The same structures resist valgus stress in extension as well. It is the lateral collateral ligament, anconeus, and joint capsule that provide stability under varus stress. Force generated at the elbow can reach up to three times body weight in everyday activities (136).

The Wrist

The wrist or carpus is a very complicated joint consisting of multiple articulations. It provides a stable support for the hand, allowing for the transmission of grip forces as well as positioning of the hand and digits for fine movements. Wrist position affects the ability of the fingers to operate (flex-extend) and to effectively grasp. The main function of the wrist is to fine-tune grasp by controlling the length–tension relationship in the extrinsic muscles to the hand (137). The hand is the principal instrument of touch and its combination of sensibility and motor function has classified it as the ultimate supportive organ of information, accomplishment, and evolution. The self-selected wrist position for maximal power grip has been shown to be 35° of extension with 7° of ulnar deviation. In addition, full wrist flexion deteriorates the efficiency of the finger flexors and the grip strength to 25% of that available when the wrist is in extension. Martial arts experts have exploited this knowledge for a long time to disarm assailants. When the wrist is positioned in flexion, the combination of a passive extensor tenodesis effect, increasing resistance to flexion, and decreasing excursion of the flexors results in a weakened grip. The assailant, therefore, is unable to maintain a grip on the weapon. The strength of the finger flexors is more than twice that of extensors. The unique feature of the metacarpophalangeal (MCP) joints is their structural asymmetry, which assisted by the ligament positions and the bony configuration of the metacarpal heads results in effective stabilization and finely tuned motion.

OVERVIEW

Human diarthroidal joints can support high levels of mechanical load for several decades, yet degenerative joint diseases occur in epidemic proportions every year. Understanding the role of mechanics in diseases, such as osteoarthritis, requires analysis of the behavior of both healthy and pathological joints under a full range of physiological loading conditions. Relationships between joint function, meniscus injury, and osteoarthritis after trauma (e.g., ACL injury–reconstruction) requires detailed knowledge of internal tissue stresses. Such knowledge is necessary to advance our understanding of cartilage–ligament–meniscus failure mechanisms at the macroscopic level, as well as mechanotransduction at the cellular level.

Experimental measurements alone are not sufficient to delineate the detailed biomechanics of the human joint. Recent publications have confirmed that mathematical modeling can be an effective tool for the simulation and analysis of complex biological structures (e.g., the human knee). However, existing modeling strategies, based on generic geometry and/or driven with simplistic kinematics and loading assumptions, have been of limited value in addressing clinical hypotheses. Research efforts in the biomechanics community have focused on cadaveric studies for the characterization of tolerances on material, geometry, and attachment parameters that will restore contact pressure distribution to within a specified deviation from accepted normal values. Models driven directly by *in vivo* kinematic data have not been possible due to the relatively poor accuracy available from traditional motion analysis methods, limiting the predictive power of FE modeling techniques in the study of knee pathogenesis. New advances in modeling problem formulations that take advantage of patient-specific geometry and unique motion analysis systems (cineradiography) for high accuracy 3D knee kinematics can overcome limitations of previous approaches. This should generate high quality estimates of internal tissue stresses, providing new insight into the role of tissue loading in the development and progression of musculoskeletal diseases such as osteoarthrosis.

BIBLIOGRAPHY

1. Maroudas A. Physical chemistry of articular cartilage and the intervertebral disc. In: Sokoloff L, ed. The Joints and Synovial Fluid. New York: Academic; 1980: 239–291.
2. Ateshian GA, Lai WM, Zhu WB, Mow VC. An asymptotic solution for the contact of two biphasic cartilage layers. J Biomech 1994;27:1347–1360.
3. Mow VC, Ratcliffe A. Structure and function of articular cartilage and meniscus. In: Mow VC, Hayes WC, eds. Basic Orthopedic Biomechanics. 2nd ed. Philadelphia: Lippincott-Raven; 1997. p 113–177.
4. Wittenburg J. Dynamics of Systems of Rigid Bodies. Stuttgart, Germany: B.G. Teubner; 1977.
5. Woltring HJ. Planar control in multi-camera calibration for 3-D gait studies. J Biomech 1980;13:39–48.
6. Woltring HJ. Representation and calculation of 3-D joint movement. Human Movement Sci 1991;10:603–616.
7. Goldstein H. Classical Mechanics. Reading (MA): Addison-Wesley; 1980.

8. Dempster WT. Space Requirements of the Seated Operator. Wright Patterson Air Force Base, OH, 1955.
9. Drillis R, Contini R, Bluestein M. Body segment parameters; A survey of measurement techniques. Artif Limbs 1964;25: 44–66.
10. Full RJ, Koditschek DE. Templates and anchors: neuromechanical hypotheses of legged locomotion on land. J Exp Biol 1999;202:3325–3332.
11. Nigg BM, Herzog W. Biomechanics of the musculoskeletal system. West Sussex (UK): Wiley; 1999.
12. Manal K, McClay Davis I, Galinat B, Stanhope S. The accuracy of estimating proximal tibial translation during natural cadence walking: bone vs. skin mounted targets. Clin Biomech (Bristol, Avon) 2003;18:126–131.
13. Tashman S, Anderst W. Skin motion artifacts at the knee during impact movements. Seventh Annu Meet, Gait and Clinical Movement Analysis Soc. Chattanooga (TN); 2002.
14. Tashman S, Anderst W. Skin motion artifacts at the knee during impact movements. Gait Posture 2002;16:11–12.
15. Alexander EJ, Andriacchi TP. Correcting for deformation in skin-based marker systems. J Biomech 2001;34:355–361.
16. Andriacchi TP, Toney MK. A point cluster method for in vivo measurement of limb segment movement. Adv Bioeng 1994; 28:185–186.
17. Fleming BC et al. The strain behavior of the anterior cruciate ligament during stair climbing: An in vivo study. Arthroscopy 1999;15:185–191.
18. Shepherd DE, Seedhom BB. Thickness of human articular cartilage in joints of the lower limb. Ann Rheum Dis 1999;58: 27–34.
19. Tashman S, Anderst W. In-vivo measurement of dynamic joint motion using high speed biplane radiography and CT: Application to canine ACL deficiency. J Biomech Eng 2003;125:238–245.
20. Jurvelin JS, Arokoski JP, Hunziker EB, Helminen HJ. Topographical variation of the elastic properties of articular cartilage in the canine knee. J Biomech 2000;33: 669–675.
21. Mow VC, Kuei SC, Lai WM, Armstrong CG. Biphasic creep and stress relaxation of articular cartilage in compression? Theory and experiments. J Biomech Eng 1980;102: 73–84.
22. Mow VC, Lai WM, Holmes MH. Advanced theoretical and experimental techniques in cartilage research. Biomechanics: Principles and Applications. Boston: Martinus Nijhoff Publishers; 1982. p 47–74.
23. Fregly BJ, Sawyer WG. Estimation of discretization errors in contact pressure measurements. J Biomech 2003;36: 609–613.
24. Harris ML, Morberg P, Bruce WJ, Walsh WR. An improved method for measuring tibiofemoral contact areas in total knee arthroplasty: A comparison of K-scan sensor and Fuji film. J Biomech 1999;32:951–958.
25. Wu JZ, Herzog W, Epstein M. Articular joint mechanics with biphasic cartilage layers under dynamic loading. J Biomech Eng 1998;120:77–84.
26. Vermilyea ME, Spilker RL. Hybrid and mixed-penalty finite-elements for 3-D analysis of soft hydrated tissue. Int J Numer Methods Eng 1993;36:4223–4243.
27. Prendergast PJ, van Driel WD, Kuiper JH. A comparison of finite element codes for the solution of biphasic poroelastic problems. Proc Inst Mech Eng [H] 1996;210:131–136.
28. Hirsch C. A contribution to the pathogenesis of chondromalacia of the patella. Acta Chir Scand 1944;83:1–106.
29. Askew MJ, Mow VC. The Biomechanical Function of the Collagen Ultrastructure of articular cartilage. J Biomech Eng 1978;100:105–115.
30. Spilker RL, Maxian TA. A mixed-penalty finite-element formulation of the linear biphasic theory for soft-tissues. Int J Numer Methods Eng 1990;30:1063–1082.
31. van der Voet A. A comparison of finite element codes for the solution of biphasic poroelastic problems. Proc Inst Mech Eng [H] 1997;211:209–211.
32. Wu JZ, Herzog W, Ronsky J. Modeling axi-symmetrical joint contact with biphasic cartilage layers–an asymptotic solution. J Biomech 1996;29:1263–1281.
33. Donzelli PS, Spilker RL, Ateshian GA, Mow VC. Contact analysis of biphasic transversely isotropic cartilage layers and correlations with tissue failure. J Biomech 1999;32: 1037–1047.
34. Wu JZ, Herzog W, Epstein M. Evaluation of the finite element software ABAQUS for biomechanical modelling of biphasic tissues. J Biomech 1998;31:165–169.
35. Blankevoort L, Kuiper JH, Huiskes R, Grootenboer HJ. Articular contact in a three-dimensional model of the knee. J Biomech 1991;24:1019–1031.
36. Eberhardt AW, Keer LM, Lewis JL, Vithoontien V. An analytical model of joint contact. J Biomech Eng 1990;112: 407–413.
37. Hirokawa S. Three-dimensional mathematical model analysis of the patellofemoral joint. J Biomech 1991;24: 659–671.
38. Oloyede A, Flachsmann R, Broom ND. The dramatic influence of loading velocity on the compressive response of articular cartilage. Connect Tissue Res 1992;27:211–224.
39. Buschmann MD et al. Confined compression of articular cartilage: linearity in ramp and sinusoidal tests and the importance of interdigitation and incomplete confinement. J Biomech 1998;31:171–178.
40. Bursac PM, Obitz TW, Eisenberg SR, Stamenovic D. Confined and unconfined stress relaxation of cartilage: appropriateness of a transversely isotropic analysis. J Biomech 1999;32:1125–1130.
41. Haut Donahue TL, Hull ML, Rashid MM, Jacobs CR. How the stiffness of meniscal attachments and meniscal material properties affect tibio-femoral contact pressure computed using a validated finite element model of the human knee joint. J Biomech 2003;36:19–34.
42. Beaugonin M, Haug E, Cesari D. Improvement of numerical ankle/foot model: Modeling of deformable bone. Proc 1997 41st Stapp Car Crash Conf. Lake Buena Vista (FL): SAE, Warrendale (PA); 1997. p 225–237.
43. Beillas P, et al. Limb: Advanced FE model and new experimental data. Stapp Car Crash J 2001;45:469–494.
44. Godest AC et al. Simulation of a knee joint replacement during a gait cycle using explicit finite element analysis. J Biomech 2002;35:267–275.
45. Halloran J, Petrella A, Rullkoetter P. Explicit finite element model predicts TKR mechanics. 49th Annu Meet, Orthopaedic Res Soc. New Orleans (LA); 2003. p 1312.
46. Benvenuti J-F. Modélisation tridimensionnelle du genou humain. Laboratoire de Génie Médical. Lausanne, Switzerland: EPF Lausanne; 1998.
47. Beillas P et al. Foot and ankle finite element modeling using CT-scan data. 43rd Stapp Car Crash Conf. San Diego; 1999. p 217.
48. Keyak JH, Skinner HB. Three-dimensional finite element modelling of bone: Effects of element size. J Biomed Eng 1992;14:483–489.
49. Ulrich D et al. The quality of trabecular bone evaluated with micro-computed tomography, FEA and mechanical testing. Stud Health Technol Inform 1997;40:97–112.
50. Ulrich D, van Rietbergen B, Weinans H, Ruegsegger P. Finite element analysis of trabecular bone structure: a comparison

of image-based meshing techniques. J Biomech 1998;31: 1187–1192.

51. Beillas P. Modélisation des membres inférieurs en situation de choc automobile. Laboratoire de Biomécanique. Paris, France: École Nationale Supérieure d'Arts et Métiers; 1999.
52. Noailles J. Modélisation mécanique par éléments finis de l'articulation du genou. Laboratoire de Biomécanique. Paris, France: École Nationale Supérieure d'Arts et Métiers; 1999.
53. Limbert GMT, Freeman MAR. Three dimensional finite element model of the human ACL. Simulation of a passive knee flexion cycle. Analysis of deformations and stresses. 47th Annu Meet Orthopaedic Res Soc. San Francisco (CA); 2001. p 794.
54. Hirokawa S, Tsuruno R. Hyper-elastic model analysis of anterior cruciate ligament. Med Eng Phys 1997;19:637–651.
55. Butler DL et al. Location-dependent variations in the material properties of the anterior cruciate ligament. J Biomech 1992;25:511–518.
56. Yamamoto K, Hirokawa S, Kawada T. Strain distribution in the ligament using photoelasticity. A direct application to the human ACL. Med Eng Phys 1998;20:161–168.
57. Pioletti DP. Viscoelastic properties of soft tissues: Application to knee ligaments and tendons. Departement de physique Ecole polytechnique federale de Lausanne. Lausanne: Ecole polytechnique federale de Lausanne; 1997.
58. Hirokawa S, Tsuruno R. Three-dimensional deformation and stress distribution in an analytical/computational model of the anterior cruciate ligament. J Biomech 2000;33: 1069–1077.
59. Taylor M, Tanner KE, Freeman MA, Yettram AL. Cancellous bone stresses surrounding the femoral component of a hip prosthesis: An elastic-plastic finite element analysis. Med Eng Phys 1995;17:544–550.
60. Aspden RM. A model for the function and failure of the meniscus. Eng Med 1985;14:119–122.
61. Anderst WJ, Tashman S. A method to estimate in vivo dynamic articular surface interaction. J Biomech 2003;36: 1291–1299.
62. Walker PS, Hajek JV. The load-bearing area in the knee joint. J Biomech 1972;5:581–589.
63. Fukubayashi T, Kurosawa H. The contact area and pressure distribution pattern of the knee. A study of normal and osteoarthrotic knee joints. Acta Orthop Scand 1980;51: 871–879.
64. Warner JJ. Articular contact patterns of the normal glenohumeral joint. J Shoulder Elbow Surg 1998;7:381–388.
65. Scherrer PK, Hillberry BM, Van Sickle DC. Determining the in vivo areas of contact in the canine shoulder. J Biomech Eng 1979;101:271–278.
66. Soslowsky LJ et al. Quantitation of in situ contact areas at the glenohumeral joint: A biomechanical study. J Orthop Res 1992;10:524–534.
67. Dennis DA, Komistek RD, Hoff WA, Gabriel SM. In vivo knee kinematics derived using an inverse perspective technique. Clin Orthop Relat Res 1996: 107–117.
68. Sheehan FT, Zajac FE, Drace JE. Using cine phase contrast magnetic resonance imaging to non-invasively study in vivo knee dynamics. J Biomech 1998;31:21–26.
69. Papaioannou G, Tashman S, Nelson F. Morphology proportional differences in the medial and lateral compartment of the distal femur. 50th Ann Meet, Orthopaedic Res Soc. San Francisco, (CA); 2004. p 1256.
70. Treece GM, Prager RW, Gee AH. Regularised marching tetrahedra: Improved iso-surface extraction. Comput Graph 1999;23:583–598.
71. Papaioannou G, Tashman S. Validation of a lower limb model based on 3D knee kinematics from a high speed biplane dynamic radiography. In: Fotiadis DI, Dassios G, Kiriaki K, Massalas CV, eds. Scattering Theory and Biomedical Engineering Modelling and Applications. World Scientific; 2001.
72. Ateshian GA. A B-spline least-squares surface-fitting method for articular surfaces of diarthrodial joints. J Biomech Eng 1993;115:366–373.
73. Paul JP, Poulson J. The analysis of forces transmitted by joints in the human body. 5th Int Conf Experimental Stress Analysis. Udine, Italy; 1974.
74. Hefzy MS, Zoghi M, Jackson WT, DiDio LJA. Method to measure the three-dimensional patello-femoral tracking. Adv Bioeng 1988;8:47–49.
75. Papaioannou G, Daly D, Spaepen A. Use of new optimalization tools in knee joint modelling. In: Dassios G, Fotiadis DI, Kiriaki K, Massalas CV, eds. Scattering Theory and Biomedical Engineering Modelling and Applications. Singapore: World Scientific; 2000. pp 282–295.
76. Jonkers I, Spaepen A, Papaioannou G, Stewart C. An EMG-based, muscle driven forward simulation of single support phase of gait. J Biomech 2002;35:609–619.
77. Papaioannou G. A Three Dimensional Mathematical Model of the Knee Joint. Bioengineering Ph.D. Thesis. Glasgow, (UK): University of Strathclyde; 2000.
78. Van der Helm FC, Veeger HE, Pronk GM, Van der Woude LH, Rozendal RH. Geometry parameters for musculoskeletal modelling of the shoulder system. J Biomech 1992; 25:129–144.
79. Zajac FE. Muscle coordination of movement: a perspective. J Biomech 1993;26(Suppl 1):109–124.
80. Pandy MG, Berme N. A numerical method for simulating the dynamics of human walking. J Biomech 1988;21:1043–1051.
81. Chao EY, Rim K. Application of optimization principles in determining the applied moments in human leg joints during gait. J Biomech 1973;6:497–510.
82. Townsend MA, Tsai TC. Biomechanics and modelling of bipedal climbing and descending. J Biomech 1976;9:227–239.
83. Chou LS, Song SM, Draganich LF. Predicting the kinematics and kinetics of gait based on the optimum trajectory of the swing limb. J Biomech 1995;28:377–385.
84. Pandy MG, Berme N. Quantitative assessment of gait determinants during single stance via a three-dimensional model—Part 2. Pathological gait. J Biomech 1989;22:725–733.
85. Pandy MG, Berme N. Quantitative assessment of gait determinants during single stance via a three-dimensional model—Part 1. Normal gait. J Biomech 1989;22:717–724.
86. Winter DA. The Biomechanics and Motor Control of Human Gait. Waterloo: University of Waterloo Press; 1987.
87. Huang CY, Stankiewicz A, Ateshian GA, Mow VC. Anisotropy, inhomogeneity, and tension-compression nonlinearity of human glenohumeral cartilage in finite deformation. J Biomech 2005;38:799–809.
88. Chan B, Donzelli PS, Spilker RL. A mixed-penalty biphasic finite element formulation incorporating viscous fluids and material interfaces. Ann Biomed Eng 2000;28:589–597.
89. Zhang H, Totterman S, Perucchio R, Lerner AL. Magnetic resonance image based 3D poroelastic finite element model of tibio-menisco-femoral contact. Proc 23rd Annu Meet Am Soc Biomechanics. Pittsburgh; 1999.
90. Perie D, Hobatho MC. In vivo determination of contact areas and pressure of the femorotibial joint using non-linear finite element analysis. Clin Biomech (Bristol, Avon) 1998;13: 394–402.
91. Eckstein F et al. Quantitative relationships of normal cartilage volumes of the human knee joint—assessment by

magnetic resonance imaging. Anat Embryol (Berlin) 1998;197: 383–390.

92. Eckstein F, Reiser M, Englmeier KH, Putz R. In vivo morphometry and functional analysis of human articular cartilage with quantitative magnetic resonance imaging—from image to data, from data to theory. Anat Embryol (Berlin) 2001;203:147–173.

93. Rudert MJ, et al. Articular cartilage thickness measurement with MRI and multi-detector computed tomography (MDCT). 49th Annu Meet, Orthopaedic Res Soc. New Orleans (Lo); 2003. p 0571.

94. Bendjaballah MZ, Shirazi-Adl A, Zukor DJ. Biomechanics of the human knee joint in compression: Reconstruction, mesh generation and finite element analysis. Knee 1995;2:69–79.

95. Xia Y. Magic-angle effect in magnetic resonance imaging of articular cartilage: A review. Invest Radiol 2000;35:602–621.

96. DiSilvestro MR, Zhu Q, Suh JK. Biphasic poroviscoelastic simulation of the unconfined compression of articular cartilage: II—Effect of variable strain rates. J Biomech Eng 2001;123:198-200.

97. Eberhardt AW, Lewis JL, Keer LM. Normal contact of elastic spheres with two elastic layers as a model of joint articulation. J Biomech Eng 1991;113:410–417.

98. Donahue TL, Hull ML, Rashid MM, Jacobs CR. A finite element model of the human knee joint for the study of tibio-femoral contact. J Biomech Eng 2002;124:273–280.

99. Weinans H, Sumner DR, Igloria R, Natarajan RN. Sensitivity of periprosthetic stress-shielding to load and the bone density-modulus relationship in subject-specific finite element models. J Biomech 2000;33:809–817.

100. Papaioannou G, Yang K, Fyhrie D, Tashman S. Validation of a subject specific finite element model of the human knee developed for in-vivo tibio-femoral contact analysis. 50th Annu Meet, Orthopaedic Res Soc. San Francisco; 2004. p 0358.

101. Couteau B et al. Finite element modelling of the vibrational behaviour of the human femur using CT-based individualized geometrical and material properties. J Biomech 1998;31:383–386.

102. Papaioannou G, Anderst W, Tashman S. Elevated joint contact forces in ACL-reconstructed knees: A finite element analysis driven by in vivo kinematic data. IMECE'03:2003 ASME Int Mech Eng Cong Exposition. Washington, DC; 2003.

103. Beillas P, Papaioannou G, Tashman S, Yang KH. A new method to investigate in vivo knee behavior using a finite element model of the lower limb. J Biomech 2004;37:1019–1030.

104. Li G, Lopez O. Reliability of a 3D finite element model constructed using magnetic resonance images of a knee for joint contact stress analysis. 23rd Proc Am Soc Biomech. Pittsburgh; 1999.

105. Charras GT, Guldberg RE. Improving the local solution accuracy of large-scale digital image-based finite element analyses. J Biomech 2000;33:255–259.

106. Dunbar WL, Jr., Un K, Donzelli PS, Spilker RL. An evaluation of three-dimensional diarthrodial joint contact using penetration data and the finite element method. J Biomech Eng 2001;123:333–340.

107. Spilker RL, Donzelli PS, Mow VC. A transversely isotropic biphasic finite element model of the meniscus. J Biomech 1992;25:1027–1045.

108. Morrison JB. The mechanics of the knee joint in relation to normal walking. J Biomech 1970;3:51–61.

109. Kettelkamp DB, Jacobs AW. Tibiofemoral contact area—determination and implications. J Bone Joint Surg Am 1972;54:349–356.

110. Radin EL et al. Relationship between lower limb dynamics and knee joint pain. J Orthop Res 1991;9:398–405.

111. Collins JJ, Whittle MW. Impulsive forces during walking and their clinical implications. Clin Biomech 1989;4:179–187.

112. Tashman S, Leisen JC, Sherlitz C, Radin EL. Methods for the reduction of heelstrike impulsive loading. Second World Cong Biomech. Amsterdam, The Netherlands; 1994.

113. Rolf C et al. An experimental in vivo method for analysis of local deformation on tibia, with simultaneous measures of ground reaction forces, lower extremity muscle activity and joint motion. Scand J Med Sci Sports 1997;7:144–151.

114. Tashman S, Anderst W. Internal/external and varus/valgus knee rotations are different in ACL-reconstructed and contralateral (intact) limbs during running. 49th Annu Meet, Orthopaedic Res Soc. New Orleans (LA); 2003. p 0124.

115. Papaioannou G, Fyhrie D, Tashman S. Effects of patient-specific cartilage geometry on contact pressure: An in-vivo finite element model Of ACL reconstruction. 50th Annu Meet, Orthopaedic Res Soc. San Francisco; 2004. p 1289.

116. Hollister SJ, Fyhrie DP, Jepsen KJ, Goldstein SA. Application of homogenization theory to the study of trabecular bone mechanics. J Biomech 1991;24:825–839.

117. Jacobs CR et al. NACOB presentation to ASB Young Scientist Award: Postdoctoral. The impact of boundary conditions and mesh size on the accuracy of cancellous bone tissue modulus determination using large-scale finite- element modeling. North American Congress on Biomechanics. J Biomech 1999;32:1159–1164.

118. van Rietbergen B et al. Tissue stresses and strain in trabeculae of a canine proximal femur can be quantified from computer reconstructions. J Biomech 1999;32:443–451.

119. Yeni YN, Fyhrie DP. Finite element calculated uniaxial apparent stiffness is a consistent predictor of uniaxial apparent strength in human vertebral cancellous bone tested with different boundary conditions. J Biomech 2001;34: 1649–1654.

120. Blankevoort L, Huiskes R, de Lange A. The envelope of passive knee joint motion. J Biomech 1988;21:705–720.

121. Ahmed AM, Burke DL. In-vitro measurement of static pressure distribution in synovial joints—Part I: Tibial surface of the knee. J Biomech Eng 1983;105:216–225.

122. Besnault B, et al. A parametric finite element of the human pelvis. 42nd Stapp Car Crash Conf Proc. Tempe (AZ); 1998. p 337.

123. Anderst W, Tashman S. A unique method to determine dynamic in vivo articular surface interaction. 4th World Cong Biomecha. Calgary, Alberta, Canada; 2002. p 520.

124. Anderst WJ, Les C, Tashman S. In vivo serial joint space measurements during dynamic loading in a canine model of osteoarthritis. Osteoarth Cartilage 2005;13(9):808–816.

125. Demetropoulos CK. Dynamic evaluation of contact pressure and effect of graft harvest at osteochondral donor site in the knee, personal communication. 2003.

126. Kempson GE, Spivey CJ, Swanson SA, Freeman MA. Patterns of cartilage stiffness on normal and degenerate human femoral heads. J Biomech 1971;4:597–609.

127. Johnston RC, Smidt GL. Hip motion measurements for selected activities of daily living. Clin Orthop Relat Res 1970;72:205–215.

128. Bergmann G, Graichen F, Rohlmann A. Is staircase walking a risk for the fixation of hip implants? J Biomech 1995;28:535–553.

129. Andriacchi TP, Mikosz RP, Hampton SJ, Galante JO. Model studies of the stiffness characteristics of the human knee joint. J Biomech 1983;16:23–29.

130. King AI. A review of biomechanical models. J Biomech Eng 1984;106:97–104.

131. Cavanagh PR et al. The relationship of static foot structure to dynamic foot function. J Biomech 1997;30:243–250.

132. Matsen FA, Fu FH, Hawkins RJ. The Shoulder: A Balance of Mobility and Stability. Rosemont: AAOS; 1992.
133. Chao EY, Morrey BF. Three-dimensional rotation of the elbow. J Biomech 1978;11:57–73.
134. Panjabi MM, Krag MH, Goel VK. A technique for measurement and description of three-dimensional six degree-of-freedom motion of a body joint with an application to the human spine. J Biomech 1981;14:447–460.
135. Yang KH, Latouf BK, King AI. Computer simulation of occupant neck response to airbag deployment in frontal impacts. J Biomech Eng 1992;114:327–331.
136. Walker PS. Human Joints and Their Artificial Replacements. Springfield (IL): Thomas; 1977.
137. Youm Y, Yoon YS. Analytical development in investigation of wrist kinematics. J Biomech 1979;12:613–621.

See also CARTILAGE AND MENISCUS, PROPERTIES OF; HIP JOINTS, ARTIFICIAL; HUMAN SPINE, BIOMECHANICS OF; LIGAMENT AND TENDON, PROPERTIES OF.

JOINT REPLACEMENT. See MATERIALS AND DESIGN FOR ORTHOPEDIC DEVICES.

LAPARASCOPIC SURGERY. See MINIMALLY INVASIVE SURGERY.

LARYNGEAL PROSTHETIC DEVICES

GUIDO BELFORTE
MASSIMILIANA CARELLO
Politecnico di Torino
Torino, Italy

GUIDO BONGIOANNINI
MAURO MAGNANO
ENT Division Mauriziano
Hospital
Torino, Italy

INTRODUCTION

The larynx is a uniquely complicated organ strategically located at the division of the upper aerodigestive tract into the gastrointestinal tract and the airways. Alteration of its function can have a significant impact on vocal, digestive, and respiratory physiology (1–6). The hyoid bone, the thyroid cartilage, and the cricoid cartilage form the outside framework of the larynx. The mobile interior framework consists of the leaf-shaped epiglottis and the arytenoid cartilages. Each vocal cord stretches from an anterior projection of the arytenoid to the anterior midline of the inside thyroid cartilage. The arytenoid cartilages move in both a rocking and sliding motion on the cricoid cartilage to abduct and adduct the true vocal cords. The intrinsic muscles of the larynx control the movement of the vocal cords.

A section of the larynx is shown in Fig. 1, which is a posterior view of the cartilages and membranes (skeleton of larynx).

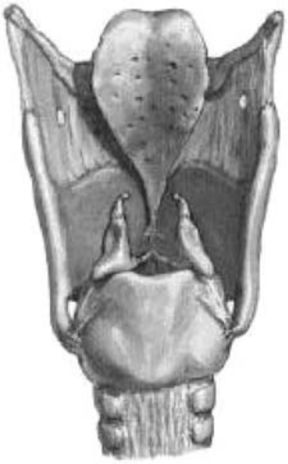


Figure 1. Section of the larynx: posterior view of cartilages and membranes.

The larynx has three important functions: protection of the lower airways during swallowing; respiration; and phonation.

Phonation is a complicated process in which sound is produced for speech. During phonation, the vocal folds are brought together near the center of the larynx by muscles attached to the arytenoids. As air is forced through the vocal folds, they vibrate and produce sound. The tone and level of the sound can be changed by contracting or relaxing the muscles of the arytenoids. As the sound produced by the larynx travels through the throat and mouth, it is further modified to produce speech.

Cancer of the larynx represented ~0.7% of the total cancer risk in 2001, and is the most common of all head and neck cancers. However, head and neck cancers account for only ~9% of all cancers diagnosed annually.

Laryngeal cancer occurs about five times more frequently in males than females. It is rare prior to age 40, after which the incidence in males increases rapidly with age. Cigarette smoking is the most important cause of laryngeal cancer, with smokers having a roughly 10-fold higher risk than nonsmokers. Heavy alcohol consumption also is a well-established risk factor.

Though laryngeal cancer is infrequent compared to cancer of the breast, lung, and prostate, the literature regarding this disease is substantial. This apparently disproportionate body of writing reflects the perceived importance of this neoplasm, which is in turn related to its potential impact on people's communicative ability: the threat to a patient's vocal organ is associated with profound psychological and socioeconomic overtones.

Originally, larynx cancer treatment focused primarily on cure by relentless and aggressive surgery. That era has been followed by the emergence of conservative partial laryngectomy, the development of more sophisticated radiation methods, and organ-sparing strategies in which chemotherapeutic, radiotherapeutic, and surgical techniques are used in a variety of combinations.

Total laryngectomy is one of the standard operations for laryngeal carcinomas. The prognosis associated with laryngeal carcinoma has improved: As the curability of laryngeal carcinoma is now $>60\%$, many patients thus survive for a long time after surgery.

Laryngectomy changes the anatomy: The lower respiratory tract is separated from the vocal tract and from the upper digestive tract, and the laryngectomee breathes through a tracheostoma. The direct connection between the vocal tract and the upper digestive tract remains unchanged.

Figure 2 shows the anatomic structure before laryngectomy (a) and after laryngectomy (b). Before laryngectomy (Fig. 2a), air can travel from the lungs to the mouth (as represented by the arrows), and the voice can be produced and modulated. After the laryngectomy (Fig. 2b), air issuing from the tracheostoma cannot reach the mouth, and sounds cannot be produced.

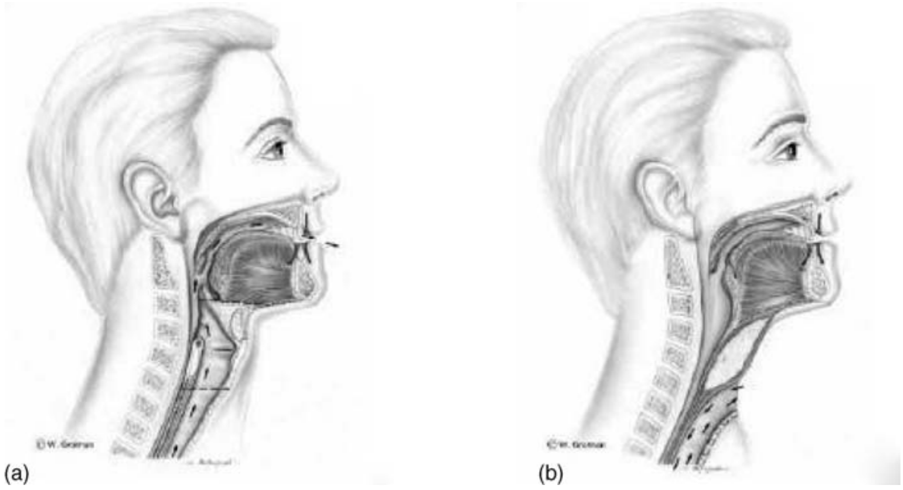


Figure 2. Anatomy before (a) and after (b) laryngectomy. (From Ref. 6.)

THE VOICE AFTER A LARYNGECTOMY

After laryngectomy, the patient is deprived of both the vibrating sound source (the vocal folds) and the energy source for voice production, as the air stream from the lungs is no longer connected to the vocal tract (1–9).

For some laryngectomy patients, the loss of speech is more important than survival itself. Consequently, a number of different methods for recovering phonation have been developed, including: esophageal speech; artificial larynx; and surgical laryngoplasty.

Numerous internal or external mechanical vibration sources have been developed that cause the air in the vocal tract to vibrate. These vibration sources can be powered by air pressure (expired air from the tracheostoma in some cases) or by an electric source with battery, and are thus classified as pneumo-larynges or electrical larynges.

ESOPHAGEAL SPEECH

Rehabilitation of the laryngectomized patient is usually a delayed process following recovery from surgery. After surgery, some patients try aphonic lip speech enhanced by buccal air trapping, while others choose written language as the method of communication.

Though most laryngectomized patients begin to learn esophageal speech, this method of speech rehabilitation requires a sequence of training sessions to develop the ability to insufflate the esophagus by inhaling or injecting air through coordinated muscle activity of the tongue, cheeks, palate, and pharynx.

Patients are encouraged to attempt esophageal sound soon after they are able to swallow food comfortably and learn to produce esophageal sound by trapping air in the mouth and forcing it into the esophagus. This produces a "burp-like" tone that can be developed into the esophageal voice.

There are various techniques for transporting air into the esophagus.

With the injection technique, the tongue forces air back into the pharynx and esophagus. This takes place in two stages, with the tongue forcing the air from the mouth back into the pharynx in the first stage, and the back of the tongue propelling the air into the esophagus in the second stage. For air to be transported into the esophagus, it is extremely important that these two stages be correctly synchronized.

With the inhalation method of esophageal speech, the patient creates a pressure in the esophagus that is lower than atmospheric pressure. As a result of this pressure difference, air will flow through the mouth past the upper segment of the esophagus into the lower esophagus. The patient will need to inhale air to be able to create a low endothoracic and esophageal pressure.

The last technique of capturing air is by swallowing air into the stomach.

Voluntary air release or "regurgitation" of small volumes vibrates the cervical esophageal inlet, hypopharyngeal mucosa, and other portions of the upper aerodigestive tract to produce a "burp-like" sound. Articulation by the lips, teeth, palate, and tongue produces intelligible speech.

Esophageal speech training is time consuming, frustrating, and sometimes ineffective.

Its main disadvantage is the low success rate in acquiring useful voice production, which varies from 26 to 55%. In addition, esophageal speech results in low-pitched (60–80 Hz) and low intensity speech, whose intelligibility is often poor. Age is the most important factor in determining success or failure: older patients are less successful in learning esophageal speech.

The airway used to create the esophageal voice is shown in Fig. 3. Direction of flow is indicated by the arrows, making it possible to distinguish between the pulmonary air used for breathing and the mouth air used for speech.

THE ELECTRONIC ARTIFICIAL LARYNX

Wright introduced the first electrolarynx in 1942. The most widely used electronic artificial larynx is the handheld transcervical device, or electrolarynx. This electrical device contains a vibrating diaphragm, which is held against the throat and activated by a button to inject vibratory energy through the skin and into the hypopharynx. By mouthing

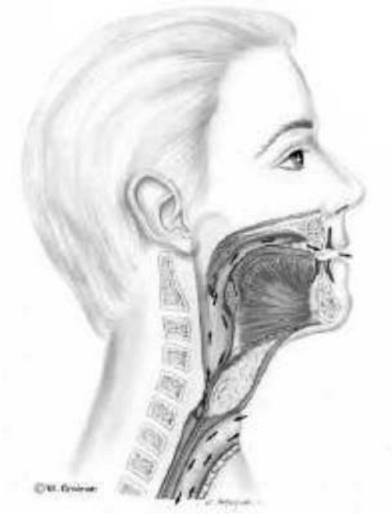


Figure 3. Oesophageal speech. (From Ref. 6.)

words, the laryngectomee converts the vibrations into a new, low frequency voice for speech.

An example of electrolarynx application is shown in Fig. 4, where it is possible to distinguish the airway from the sound way and the positioning of the device in contact with the neck. Examples of commercial electrolarynges are shown in Fig. 5.

The same operating principle has been used in recent years to develop a transoral device (Fig. 6), which is placed in the mouth, where it is housed in an upper denture or an orthodontic retainer. The system consists of a control circuit, a loud speaker, and rechargeable batteries positioned inside the denture, as well as a charging port so that the batteries can be recharged outside the mouth.

FROM THE PNEUMOLARYNX TO THE VOICE PROSTHESIS

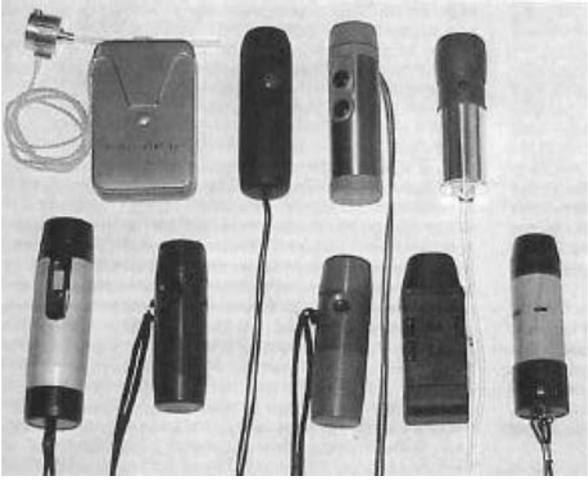


Figure 5. Examples of commercial electrolarynx. (From Ref. 5.)

The artificial larynx has undergone many transformations over the years, and continues to do so today. The first types of pneumolarynges, which included neck or mouth types and internal or external types were developed in 1860, when the first attempts at voice rehabilitation through surgery or prosthetization were made. A device was used to direct air from the lungs via a small tube to the mouth (1,3,5,7,9).

Experimental research started in 1860 with Ozermack and Burns, though it was not until 1874 that Billroth used an internal prosthesis designed by Gussenbauer (Fig. 7). This device was fitted in the trachea via the stoma with a cannula featuring a branch that entered the oral cavity.

Other similar prostheses were developed and used (Gosten, Gluck, etc.). Results, however, were not particularly satisfactory, as these devices were plagued by problems, such as tissue necrosis and leakage of food and liquids into the windpipe and lungs, thus causing infections (i.e., pneumonia).

Figure 8 shows the external prosthesis developed by Caselli in 1876. Figure 9 shows the application of the external prosthesis developed by Briani in 1950.

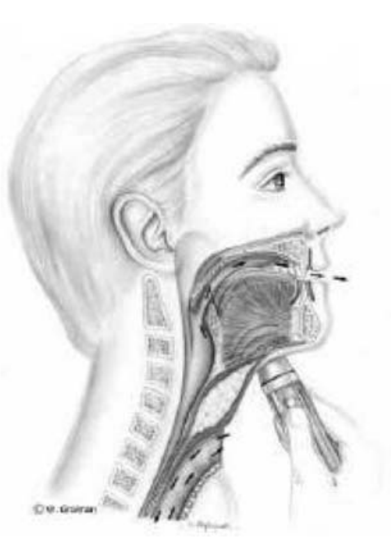


Figure 4. Electrolarynx speech. (From Ref. 6.)

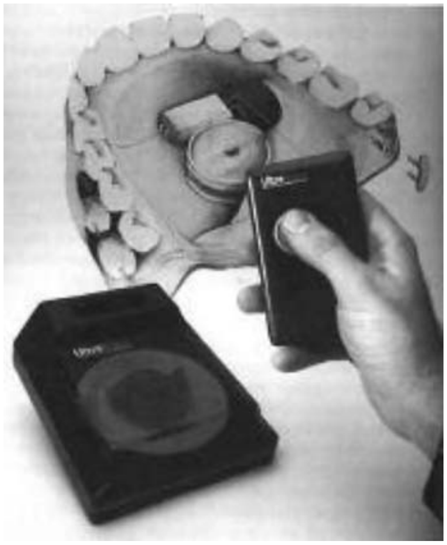


Figure 6. Transoral device. (From Ref. 5.)

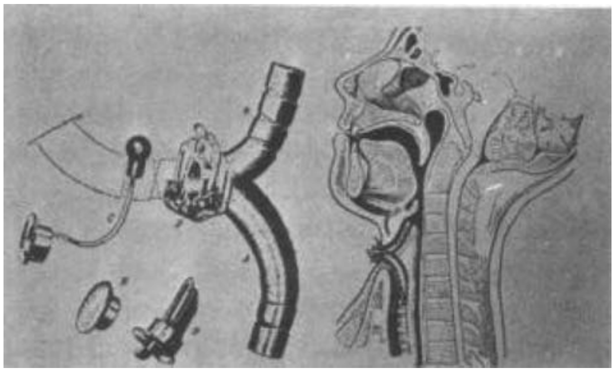


Figure 7. The Gussembauer prosthesis. (From Ref. 3.)

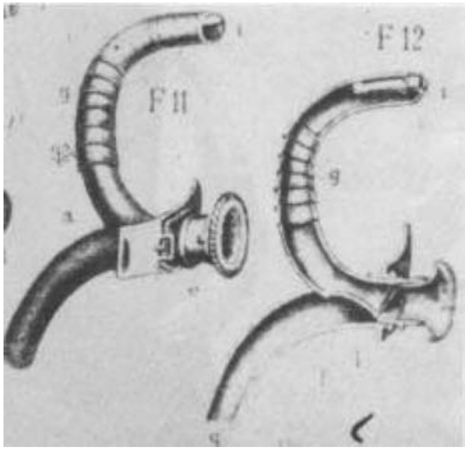


Figure 8. The Caselli prosthesis. (From Ref. 3.)

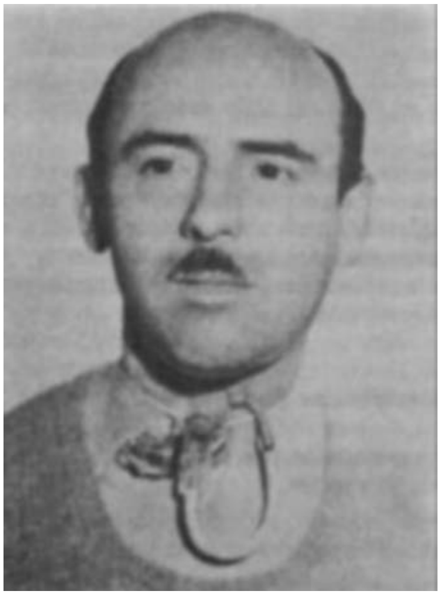


Figure 9. The Briani prosthesis. (From Ref. 3.)

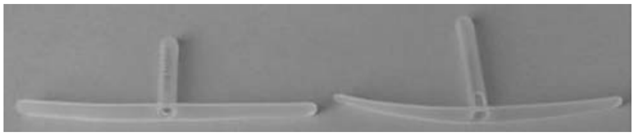


Figure 10. Example of removable prosthesis (Bivona).

Since 1960, surgical technique has been improved and a number of different types of prosthesis have been invented that take the importance of reproducing a voice that is similar to the original voice into account.

Tracheoesophageal puncture (TEP) or fistula with a voice prosthesis is the most popular of the new techniques, and has become the standard method of voice restoration after laryngectomy. However, it is usable only for selected patients.

Singer and Blom introduced the first TEP in 1980 (7,8). The aim was to create a permanent fistula (puncture or shunt) into the posterior tracheoesophageal wall between the trachea and the esophagus tract, so the pulmonary air can shunt from the airway into and up the esophagus. In this way, the vibration of the anatomic structures produces a noise.

A removable or fixed one-way type prosthesis can be positioned in the fistula.

The removable, or nonindwelling, prosthesis (Bivona, Blom-Singer duckbill) can be removed daily for cleaning by the patient, and is taped to the peritracheostomal skin. Figure 10 shows a photo of two Bivona valves; lengths differ in order to adapt the prosthesis to the patient.

The fixed, or indwelling, prosthesis cannot be removed daily, but is surgically placed in the fistula under local anesthesia. The operating principle of this type of prosthesis (known as a phonatory valve or voice button) can be explained with reference to Fig. 11.

Pulmonary air can be pushed through the valve into the esophagus for speech during expiration in two ways: by the laryngectomee, who covers the tracheal stoma with a finger (bottom left, Fig. 11) or automatically by a tracheostoma breathing valve (bottom right, Fig. 11).

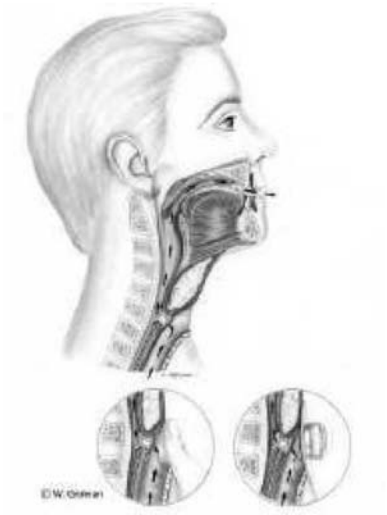


Figure 11. Speech with phonatory prosthesis. (From Ref. 6.)

The tracheostoma breathing valve contains an elastic diaphragm installed in a peristomal housing, which permits normal respiration during silent periods. Expiratory air for speech shuts off the pressure-sensitive diaphragm, and is thus diverted through the valve into the esophagus. This device eliminates the need for manual stoma occlusion during speech.

In both cases, air from the lung reaches the esophagus and causes the mucosal tissue tovibrate. The resulting sound can be modulated by the mouth, teeth, oral cavity, and so on, to produce the new voice.

Most speakers that have prosthesis do not have difficulties with articulation, rate of speech, or phonatory duration.

If esophageal speech depends on gulping or trapping air using the phonatory valve, the resulting speech depends on expiratory capacity. Voice quality is very good, and may resemble the "original" voice.

Poor digital occlusion of the tracheostoma as well as poor tracheostoma valve adherence allows pulmonary air to escape from the stoma prior to its diversion through the prosthesis into the oesophagus for voice production. In fact, the bleeding off of pulmonary air limits phonatory duration and, therefore, the number of words that can be spoken.

The one-way valve design of the prosthesis prevents aspiration (of food and liquid) from the esophagus to the trachea. An example of a phonatory valve is shown in Fig. 12 (11,12). The prosthesis illustrated in this sketch consists of an air-flow tracheal entry, whereby an endotracheal retention collar is connected to the mucosa (or tracheal flange), a hollow cylindrical tube (whose length depends on the patient's physical characteristics) connecting the trachea to the esophagus, an endoesophageal flange, and a dome (or hat) that closes the proximal endoesophageal end of the tube. Via the razor-thin slit (or esophagus exit), the hat enables airflow to pass from the trachea to the esophagus when there is a positive differential pressure, and prevents the reverse flow of liquid (or food) when the differential pressure becomes negative. The arrows represent the airflow paths.

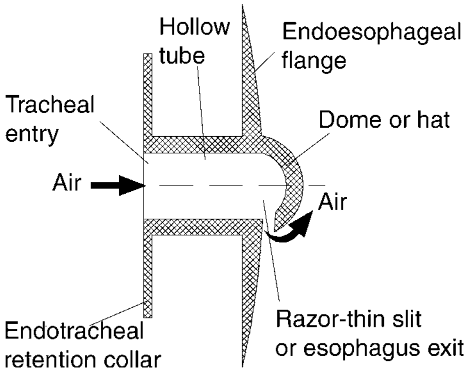


Figure 12. Sketch of phonatory valve or prostheses.

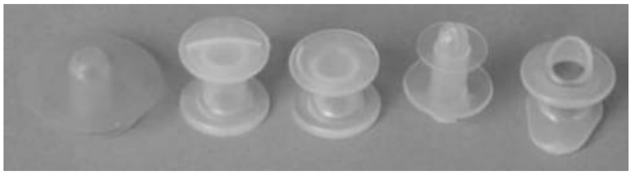


Figure 13. Examples of fixed commercial prosthesis. (Staffieri, Groningen standard, Groningen low pressure, Panje, Provox).

Hat shape and the extension of the razor-thin slit can differ according to valve type. The razor-thin slit may be located at the base of the hat (Staffieri, Groningen low pressure), at the center of the hat (Panje, Groningen standard), or inside the hollow tube (Provox, Blom-Singer).

Though valve geometry and shape may vary, the operating principle remains the same.

Several commercial prostheses are shown in Fig. 13: from left to right, they include the Staffieri, Groningen standard, Groningen low pressure, Panje, and Provox types.

Fixed and removable prostheses are available in different lengths, which usually range from 6 to 12 mm to enable the surgeon to select the dimensions that are best suited to the patient's physical characteristics, for example, posterior tracheoesophageal wall thickness.

To compare valve performance in the laboratory, most authors use valve airflow resistance (7,11), which is defined as the ratio of pressure to flow-rate. Figure 14 show an example of resistance versus flow-rate characteristics obtained with experimental tests on commercial valves (11).

Low resistance allows air to pass freely though the prosthesis with little effort on the part of the patient, and

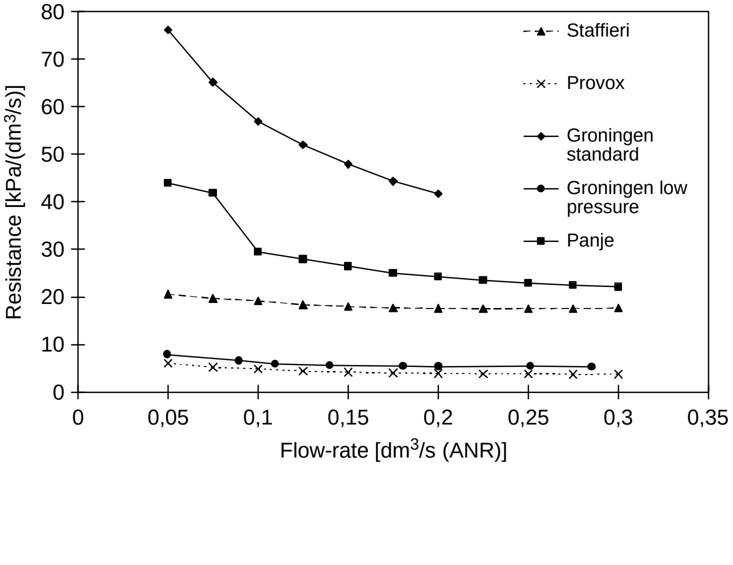


Figure 14. Resistance of commercial valves.

is thus the most favorable condition. Different materials have been used for phonatory prostheses. For indwelling prostheses, silicone rubber is the material of choice, though other materials such as polyurethane have also been used. The most significant problem affecting voice prostheses is the formation of a thick biofilm on the esophageal surface, as the esophageal environmental around the prosthesis provides ideal growing conditions for bacteria, fungi, and yeast.

In this area, secretions from the trachea, the mucous membranes in the oropharynx and esophagus, and saliva from the oral cavity create an optimal "pabulum" for microorganisms, which can thus adhere to the prosthesis surface. Though biofilm does not lead immediately to valve, malfunction, colonies growing around the prosthesis can increase airflow resistance, blocking the valve or preventing it from closing correctly. This causes saliva to leak from the oral cavity through the esophagus lumen. The biofilm also grows into the valve material (silicone rubber).

Massive microorganism growth is more frequent in indwelling prostheses than in non-indwelling types, as the latter are regularly removed and cleaned.

The first step of this colonization is the formation of a very thin layer of organic origin, called a conditioning film. Microorganisms adhere to the thin layer of saliva conditioning film on the inner esophageal surface of the device and form the first stratum of biofilm, which is nearly impossible to eradicate. There are few methods for reducing the growth of bacteria and fungi: One possibility is to apply (through oral instillations) antimycotic drugs that reduce biofilm formation and increase prosthesis life (1,2,7,8,13,14). The valve has a limited life, varying from ~4 to 8 months. After this period, the valve must be removed and changed in an outpatient procedure because of a dysfunctional mechanism caused by the biofilm, which causes food and liquids to leak from the esophageal area to the trachea.

Figure 15 shows a new Provox valve (left) and a Provox valve after eight months of use by a patient (right). Silicon rubber deterioration caused by microbial action and adherent deposits of microorganisms is clearly apparent.

Current work with voice prostheses focuses on improving their aerodynamic characteristics (low airflow resistance) and developing more resistant materials that can extend the life of the device. For the patient, in any case, the most important thing after a total laryngectomy is to be hopeful: a prosthesis can provide a new voice, improving the quality of life.

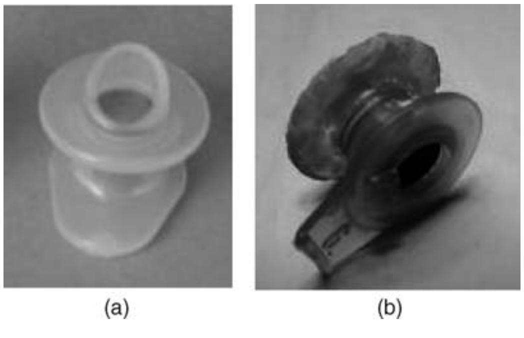


Figure 15. Provox valve: (a) new and (b) old.

BIBLIOGRAPHY

1. Mahieu HF. Voice and speech rehabilitation following laryngectomy. Ph.D. dissertation, Rijksuniversiteit Groningen; 1988.
2. Pighi GP, Cristoferi V. Protesi tacheo-esofagee. Bologna: Arianna Editrice; 1998.
3. Societé Francaise d'Oto-Rhino-Laryngologie et de Pathologie cervico-faciale. Réhabilitation de la voix et de la déglutition après chirurgie du larynx. Arnette, Paris; 1992.
4. Testut L, Jacob O. Trattato di anatomia topografica. Vol. II Collo - Torace –Addome, Utet, Torino, 1998.
5. Harrison LB, Sessions RB, Hong WK. Head and neck cancer. Lippinocott Williams & Wilkins; 2004; 178–189.
6. http://www.orl.nl
7. Blom ED. Tracheoesophageal voice restoration: origin, evolution, state of the art. Folia Phonatrica Logopaedica 2000; 52:14–23.
8. Blom ED, Singer MI, Hamaker RC. Tracheoesophageal voice restoration following total laryngectomy. San Diego: Singular Publishing Group Inc.; 1998.
9. Brown DH, et al. Postlaryngectomy voice rehabilitation: state of the art at the millennium. World J Surg 2003;27(7):824–831.
10. Nakamura T, Shimizu Y. Thacheal, laryngeal and esophageal replacement device. The Biomedical Engineering Handbook CRC and IEEE Press; 2000, p 136/1–136/13.
11. Belforte G, Carello M, Miani C, Staffieri A. Staffieri tracheo-oesophageal prosthesis for voice rehabilitation after laryngectomy: an evaluation of characteristics. Med Biol Eng Comput 1998;36:754–760.
12. Staffieri M, Staffieri A. A new voice button for post-total laryngectomy speech rehabilitation. Laryngoscope 1988; 98(9):1027–1029.
13. Schwandt LQ, et al. Prevention of biofilm formation by dairy products and N-acetylcysteine on voice prostheses in an artificial throat. Acta Otolaryngol 2004;124:726–731.
14. Leunisse C, et al. Biofilm formation and design features of indwelling silicone rubber tracheoesophageal voice prostheses—An electron microscopical study. J Biomed Mater Res 2001;58(5):556–563.

See also COMMUNICATION DEVICES; PULMONARY PHYSIOLOGY.

LASER SURGERY. See ELECTROSURGICAL UNIT (ESU).

LASERS, IN MEDICINE. See FIBER OPTICS IN MEDICINE.

LENSES, CONTACT. See CONTACT LENSES.

LENSES, INTRAOCULAR

HABIB HAMAM
Université de Moncton
Moncton, New Brunswick,
Canada

INTRODUCTION

Ridley's implantation (1949) of the first intraocular lens (IOL) marked the beginning of a major change in the practice of ophthalmology. The IOLs are microlenses placed inside the human eye to correct cataracts, nearsightedness, farsightedness, astigmatism, or presbyopia. There are two types of IOLs: anterior chamber lenses, which are placed in the anterior chamber of the eye between the iris and the cornea, and posterior chamber IOLs, which are placed in the posterior chamber behind the iris and rest against the capsular bag. Procedures for implanting the IOLs and technologies for manufacturing them in various sizes, thicknesses, and forms as well as with various materials progressed tremendously in the last decade. Multifocal IOLs are one of the important signs of this progress. While monofocal IOLs, the most commonly used, are designed to provide clear vision at one focal distance, the design of multiple optic (multifocal) IOLs aims to allow good vision at a range of distances.

INTRAOCULAR LENSES: WHAT AND WHY?

An intraocular lens, commonly called IOL, is a tiny artificial lens implanted in the eye. It usually replaces the faulty (cataractous) cristalline lens. The most common defect of the natural lens is the cataract, when this optical element becomes clouded over. Prior to the development of IOLs, cataract patients were forced to wear thick coke bottle glasses or contact lenses after the surgery. They were essentially blind without their glasses. In addition to IOLs replacing the crystalline lenses, a new family of IOLs, generally referred to as phakic lenses, is nowadays subject of active research and development. (Phakos is the Greek word for lens. Phakic is the medical term for individuals who have a natural crystalline lens. In Phakic IOL surgery, an intraocular lens is inserted into the eye without removal of the natural crystalline lens.) These IOLs are placed in the eye without removing the natural lens, as is completed in cataract surgery. They are used to correct high levels of nearsightedness (myopia) or farsightedness (hyperopia).

An IOL usually consists of a plastic lens with plastic side struts called haptics to hold the lens in place within the capsular bag. The insertion of the IOL can be done under local anesthesia with the patient awake throughout the operation, which usually takes <30 min in the hands of an experienced ophthalmologic surgeon (Fig. 1).

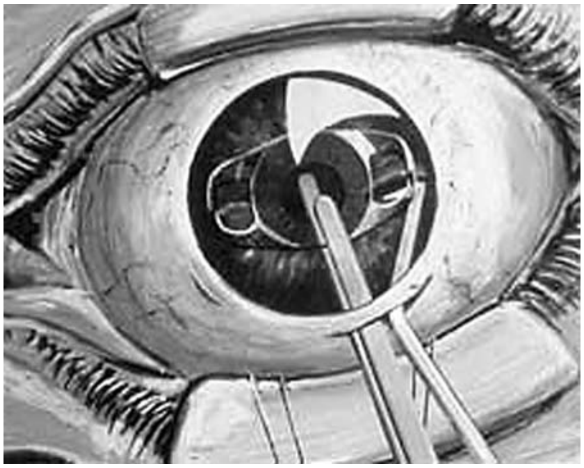


Figure 1. Implantation of an IOL: Since the natural lens is left undistrubed, the operation is much simpler than a cataract operation. The entire procedure consists of making a small incision at the edge of the cornea and placing the appropriate tiny plastic lens in the space between the iris and the cornea, (the anterior chamber). Stitches are used to close the incision.

HISTORICAL OVERVIEW

The idea of the IOL dates back to the beginning of modern cataract surgery when Barraquer developed keratomileusis (1). However, the first implantation of an artificial lens in the eye was probably attempted in 1795 (2). References to the idea of the IOL before World War II in ophthalmic literature are rare. There has been mention of limited animal experiments using both quartz and plastic material performed in the 1940s, but nothing had come of these efforts (3).

The most important step toward the implantation of IOLs came as a result of World War II pilots, and the injuries sustained when bullets would strike the plastic canopy of their aircraft (Fig. 2), causing small shards of plastic to go into their eye. In the late 1940s, Howard Ridley was an RAF ophthalmologist looking after these unfortunate pilots and observed, to his amazement, little or no reaction in cases in which the material had come from Spitfire planes. He then concluded the poly(methyl methacrylate) (PMMA) material of the canopies (windshield) was compatible with eye tissue (4). This observation sparked the idea for inserting an artificial lens in the eye. Ridley, who was convinced this lens should be placed in the posterior chamber, designed a disk-shaped lens, much like the natural lens, with a small peripheral flange allowing him to hold the lens with forceps (4). The artificial lens, made entirely of PMMA, weighed slightly > 100 mg in air and was ~8.35 mm in diameter. In several cases, he attempted to implant the lens following intracapsular surgery using the vitreous base for support (5). On November 29, 1949, the first successful IOL implantation was performed at St. Thomas Hospital in London (6,7). While far from perfect, the procedure worked well enough to encourage

Figure 2. Invention of the IOL: During World War II, Fighter pilots were sometimes involved in accidents where the plastic windshield (canopy) of their aircraft was shattered. Doctors found that fragments of the canopy that entered the eye were tolerated by the eye tissues. They might remain inside the eye for years, and the eye would not react.

further refinement. Then, over a decade, Ridley implanted several hundred IOLs (8).

Though Ridley was ahead of his time, his method was subject to serious criticism. Complications were common and failure rates > 50% were often contemporaneously quoted. Fixation was dependent on the formation of adhesions between the iris and the capsule. Several ophthalmologists strongly opposed to his procedure. Implantation in the anterior chamber was technically easier and was compatible with intracapsular surgery. Also, fixation could be achieved at the time of implantation by adding haptic struts to the lens that could be wedged into the angle. The first anterior chamber lens was implanted by Baron in 1952 (8).

To make intraocular lens implantation safe, developments in lens design and surgical techniques were required. Lens implantation did not become widely adopted as an integral part of cataract surgery until the 1980s. Key advances were the introduction of viscoelastic fluids to protect the cornea during implantation and flexible haptics to enhance long term stability of the IOL (9).

With traditional single vision monofocal IOLs, people generally experience good vision after surgery at a single focal point, either near or at a distance. The multifocal IOL (10) was designed in the mid-1990s to provide a full range of vision with independence from glasses in most situations.

Besides, the invention of phakic lenses is no less important than Ridley's invention. These IOLs were introduced by Strampelli (11) and later popularized by Barraquer in the late 1950s (12). Phakic IOLs are becoming more popular because of their good refractive and visual results and because they are easy to implant in most cases (13). In the beginning, the design was a biconcave angle-supported lens. These lenses were abandoned following serious angle- and endothelium-related complications. By the late 1980s, Baikoff (14,15) introduced a myopia lens with Kelman-type haptics (16). This design had many problems, leading to its design modification a number of times. Fyodorov, the inventor of radial keratotomy (17), introduced the concept of a soft phakic lens in the space between the iris and the anterior surface of the crystalline lens (18).

Based on earlier works of Worst, winner of the Binkhorst Award for innovation in ophthalmology in 1976, Fechner introduced phakic myopia lens of iris claw design in 1986 (19). This IOL is then referred to as Worst–Fechner lens (20). Many companies around the world manufactured it in various models. Today, people usually identify it by the name of Artisan IOL.

MATERIAL OF THE IOL

Many factors, such as the optical quality of the system of the eye (aberrations, . . .), presence of inflammation, cost, and wound size, depend on the material and the form of the IOL.

From the point of view of flexibility, there are two families of IOLs: foldable and inflexible lenses. Foldable IOLs are generally made of acrylic or silicone. They can be rolled up and inserted through a tube with a very small incision not requiring any stitches. Inflexible IOLs, typically made of PMMA, require a larger incision because they are unfoldable. Most lenses are biconvex, thus optically equivalent upside down. However, most lenses have haptics which are generally angled to push the posterior optics.

Four basic materials are used for IOLs: PMMA, silicone, acrylic and collamer. Other materials are also used. For example, some manufacturers replace silicon by a hydrophilic biocompatible polymer, called collamer. Many IOLs have been made from PMMA plastic, the same plastic the original hard contact lenses were made of. Silicon IOLs are foldable. Folding an IOL allows it to be inserted through a smaller incision. A smaller incision heals faster and induces less postop astigmatism. Some foldable lenses are now being made of acrylic plastic. While acrylic and silicone lenses are very popular, PMMA is the time-tested material but, as stated above, requires a large incision. Silicone oil is also a problem for silicone IOLs in that the view of the fundus can be severely degraded. This is less so for PMMA and hydrophobic acrylic IOLs and least for hydrophilic acrylic. Although this is a relative contraindication for silicone IOLs in the face of significant vitreoretinopathy, a solvent exists that eliminates the problem (21). Collamer is a new hydrophilic material just recently released that has shown some interesting properties. It has been shown to exhibit less internal reflectance than other lens materials including silicone, acrylic, and PMMA (22). It reduces the risk of long-term inflammation (23). Table 1 summarizes the characteristics of the four materials.

VARIOUS PARAMETERS FOR IOLs

Are age, race, and sex important parameters for IOL implantation? Age at which surgery is performed turned out to be of great importance (24–28). The ideal age should be at around 18 years when the refraction stabilizes. However, in specific circumstances, in the interest of the minor patient, the parents and the surgeon can opt to perform phakic lens implantation at an earlier age. Studies

Table 1. Four Commonly Used IOLs Materials and their Advantages and Drawbacks

Material	Flexibility	Advantages	Drawbacks
PMMA	Rigid	Low cost, less inflammation, long-term experience, good bicompatibility	larger incision, not foldable
Silicone	Foldable	Smaller incision, injectable	high cost, more inflammation, cannot use with silicon oil
Collamer	Foldable	Smaller incision, less inflammation, very good bicompatibility	high cost, short term experience
Acrylic	Foldable	Smaller incision, less inflammation, high refraction index (thin IOL), good bicompatibility	high cost

of the suitable age of IOL implantation in children have been carried out (24). A 3-year-old child has been qualified with IOL implantation, the child younger than 9 years old should be implanted with a normal adult IOL and then corrected with glasses, and a child after 10 years old should be directly implanted with a proper dioptric IOL (24). Some researchers evaluated the influence of cataract surgery on the eyes of children between 1 and 5 years old. They concluded that cataract surgery, either extraction with or without IOL implantation, did not retard axial elongation in children above 1 year old (25). Comparisons between children with congenital or developmental lens opacities who underwent extracapsular cataract extraction and children with normal eyes have been carried out (26). The pattern of axial elongation and corneal flattening was similar in the congenital and developmental groups to that observed in normal eyes. No significant retardation or acceleration of axial growth was found in the eyes implanted with IOLs compared with normal eyes. A myopic shift was seen particularly in eyes operated on at 4–8 weeks of age and it is recommended that these eyes are made 6 D hyperopic initially with the residual refractive error being corrected with spectacles (26).

To our knowledge, IOL implantation does not depend on race and sex.

OPTICAL QUALITY

Two optical qualities are distinguinshed: the intrinsic optical quality of the IOL and the optical quality of the system of the eye including the IOL. Many factors, such as the material and the geometrical profile of the IOL, influence the intrinsic quality of this optical element. Axial shift (decentration), transversal rotation (not around the optical axis), and deformation (mechanical stresses, ...) of the IOL are examples of factors affecting the optical quality of the whole system of the eye even in case the IOL alone is a perfect optical element. Thus, the optical quality of the IOL is certainly important. However, for vision, the determinant factor is the optical quality of the whole system of the eye in which the IOL is implanted. Several studies have been undertaken to assess the optical quality of the IOL and the optical quality of the whole system of the eye. Before progressing in this section, let us briefly introduce the notion of optical quality.

Aberrations

Stated in wave optics, the system of the eye should transform the input wavefront into a perfect convergent spherical wavefront that has the image point as center (29–31). Note that an optical wavefront represents a continuous surface composed of points of equal phase. Thus all image-forming rays, which travel normal to the exit spherical wavefront, meet in the focal point in phase, resulting in maximum radiant energy being delivered to that point. In reality, this situation never occurs. The rays modified by the optical system do not converge entirely to a common point image. For one object point correspond several image points that form a blurred image. This deviation from the ideal case is called optical aberration, or merely aberration, and is a measure of the optical quality of the system. Aberration can be quantified either with respect to the expected image point or to the wavefront corresponding to this ideal point. If the real output wavefront is compared to the ideal one, it is called the difference between them wavefront aberration (29). All human eyes suffer from optical aberrations that limit the quality of the retinal image (32–36). Several metrics have been proposed to measure the optical quality of the system of the eye (37–41). Let us return back to IOLs now. Optical quality of multifocal IOLs will be treated in the section devoted to this kind of lens.

Optical Quality of the IOL

The optical quality of IOL was the subject of intensive studies. Several common but some contrasted results have been obtained. An exhaustive study goes beyond the scope of this document. We limit our attention to some recent results. Tognetto et al. (42) evaluated the optical quality of different IOLs by using an optical test bench. The purpose of the study was to evaluate the optical quality of IOLs and not to evaluate the optical performance of these lenses once implanted. Three randomly acquired samples of 24 different models of foldable IOLs were compared. The conclusion is that different IOLs can transmit different spectra of spatial frequencies. The best frequency response was provided by acrylic IOLs, particularly those with an asymmetrically biconvex profile. This could be due to a reduction of optical degradation provided by this type of profile. A lens with a higher frequency response should create a better quality of vision once implanted, and the frequency response should therefore be considered when choosing the intraocular lens model (42).

Negishi et al. (43) evaluated the effect of chromatic aberrations in pseudophakic eyes with various types of IOLs. Their results show that longitudinal chromatic aberrations of some IOLs may degrade the quality of the retinal image. They concluded that attention must be paid to the detailed optical performance of IOL materials to achieve good visual function.

In a comparative study (44), Martin found that the collamer IOL reduced the number of induced higher order aberrations when compared with acrylic and silicone lenses. Indeed, he found that the collamer IOL has 55–117% fewer induced higher order aberrations than acrylic or silicone materials. As a consequence, it produces less postop glare. He concluded the collamer lens provides clearer vision than the other lenses.

Optical Quality of ARTISAN Lenses

Brunette et al. (45) evaluated the optical quality of the eye before and after the insertion of an ARTISAN phakic intraocular lens for the treatment of high myopia (range −20.50 to −9.75D). Consecutive patients implanted with the ARTISAN lens by a single surgeon were prospectively enrolled. One eye per subject was tested. The wavefront aberration was calculated from images recorded with a Hartmann-Shack sensor (46,47). The PSF and the MTF were also computed from the wavefront aberration. It was concluded that preliminary data using the Hartmann–Shack wavefront sensor have not revealed a tendency toward deterioration of the optical performance following the insertion of an ARTISAN lens for the treatment of high myopia. The Hartmann–Shack sensor is a useful tool for the objective assessment of the image optical quality of eyes with a phakic intraocular lens.

MULTIFOCAL IOLs

Unlike the natural lens, the curvature of current intraocular lenses cannot be changed by the eye. Standard intraocular lenses provide good distance vision, and the patient needs reading glasses for near vision. Newer bifocal intraocular lenses give distance vision in one area and near vision in another area of the vision field. How does it work?

The basic idea consists in providing a lens with two posterior focal points instead of one. The IOL is no longer monofocal. It becomes bifocal. Two solutions are possible: refractive or diffractive bifocal lens. A third solution consists in combining both approachs together.

Diffractive Lenses

The idea comes from the principle of the Fresnel zone plate. It consists in designing a binary diffractive phase element so when the incident wave comes across this diffractive element, all the resulting waves, coming from all points of the zone plate, arrive in phase at a certain (focal) point. They then superimpose constructively, yielding a focusing behavior. As shown in Fig. 3, waves traveling along various segments arrive in phase at the focal point since the optical paths differ by a multiple of the wavelength. To fulfill this condition, the thickness d is chosen so that it introduces a phase shift of π: $d = (2k+1)\,\lambda/[2(n-1)]$ (optical path: $(2k+1)\,\lambda/2$). In Fig. 3, $k = 0$, yielding $d = \lambda/[2(n-1)]$, where n is the refraction index of the diffractive lens. The radii of the rings (Fig. 3) verify the following rule: $r_m = \sqrt{m}\cdot r_1 = \sqrt{m\lambda f}$, where r_1 is the radius of the smallest circle and f is the focal length. Without grooves on the diffractive, the waves traveling along the segments, represented by dashed lines in Fig. 3, would arrive in phase opposition with respect to other waves. In reality, the binary form (two phase levels) of the diffractive lens does not fully ensure the condition of coming in phase at the focal point. For rigor, several phase levels are required (Fig. 4). In general, the diffraction efficiency η, which is defined as the ratio of the focused energy to the incident energy, increases with the number of phase levels L according to the following formula (48): $\eta = \sin^2(\pi/L)/(\pi/L)^2$. Using two phase levels (binary), only 41% of the energy is focused. The rest is scattered in space. A four level diffractive lens focuses 81% of the input energy (it scatters only 19% of the incident energy).

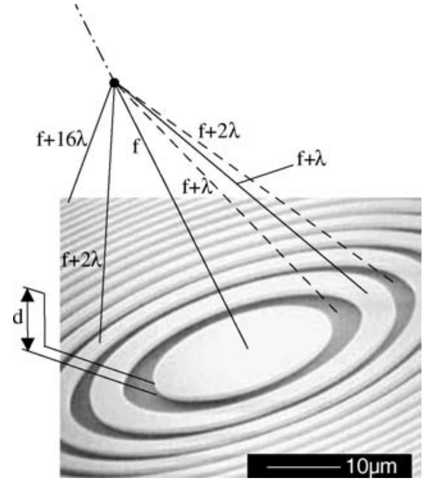


Figure 3. Binary diffractive phase element: Fresnel zone plate. Waves traveling along various segments arrive in phase at the focal point since the optical paths differ by a multiple of the wavelength. To fulfill this condition, the thickness d is chosen so that it introduces a phase shift of π: $d = (2k + 1)\ \lambda/[2(n-1)]$ (optical path: $(2k + 1)\ \lambda/2$): In the figure, $k = 0$.

To obtain a bifocal diffractive lens, we need to focus rays on two focal points at distances f_1 and f_2. It can be done by providing two series of zones (rings). The first series, the inner one, involves radii verifying $r_m^{(1)} = \sqrt{m\lambda f_1}$ (with $m = 1,2,\ldots, M$), whereas the radii in the second series, the outer one, satisfy the condition $r_p^{(2)} = \sqrt{p\lambda f_2}$ (with $p = M+1,\ldots, P$). To obtain a multifocal diffractive lens, we need more series of radii.

Refractive Lenses

An alternative to obtain a multifocal length consists in modifying the surface profil of a conventional biconvex lens so that it provides two or more different focal points for light convergence. The technique consists in designing a spherical refractive surface that has additional refracting surfaces to give a near add (Fig. 5), or a near and intermediate add. The principle of multifocal refractive lenses is illustrated in Fig. 6a. Refractive IOLs with several focal points are commercialized in various models. For example, one of the models includes five refractive zones targeting distance, intermediate and near vision (Fig. 6b). The IOL uses continuous aspheric optics to ensure that 100% of the light entering the eye reaches the retina. The lens uses five concentric zones with the first, third, and fifth zones being far dominant and second and fourth zones being near dominant (49). The light distribution is arranged so that 50% of light is distant focussed, 13% is focussed for intermediate vision and 37% for near vision. The near add

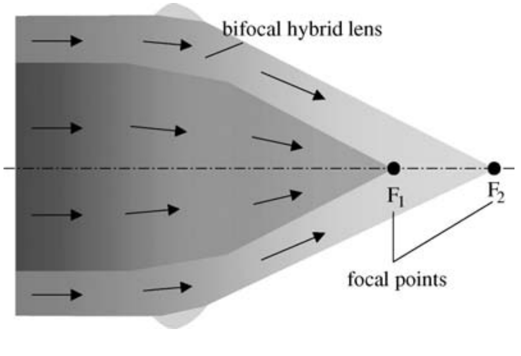


Figure 4. Bifocal hybrid refractive/diffractive IOL: Anterior surface broken up into a refractive zone and a second zone composed of concentric diffractive rings.

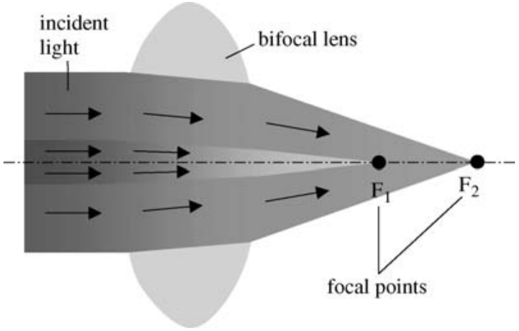


Figure 5. Bifocal refractive IOL: Spherical refractive surface that has additional refracting surfaces to give a near add.

comprises of 3.5 dioptre intraocular power equivalent to a 2.75–2.85 add in the spectacle plane (49).

Optical Quality of Refractive and Diffractive Lenses

This discussion will be limited to some recent results. Pieh et al. (50) compared the optical properties of bifocal diffractive and multifocal refractive intraocular lenses. The IOLs were manufactured by different companies. A model eye with a pupil 4.5 mm in diameter was used to determine the point spread function (PSF) (30,51) of the distance focus and near focus of the IOLs to compare them with PSFs of foci of corresponding monofocal lenses. For interpreting the PSFs the through focus response, the modulation transfer function (MTF) (51), and the Strehl ratio (51) were evaluated. They concluded the modulation transfer functions reveal comparable properties for distance vision and a superiority of the bifocal diffractive lens over the refractive multifocal lens for near vision. This mean due to the fact that the incoming light is distributed over different zones in the refractive lenses.

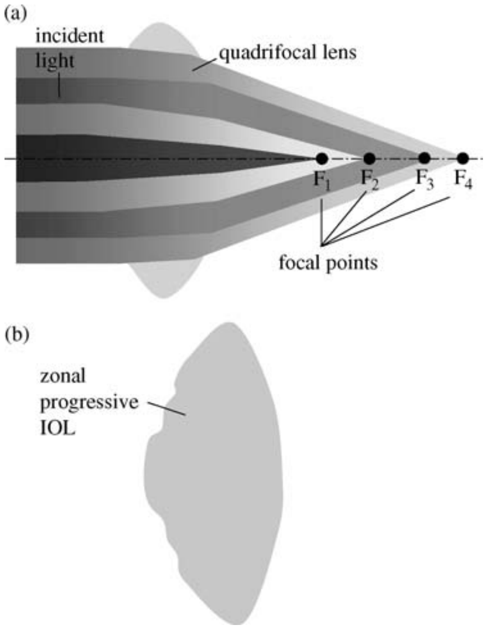


Figure 6. Multifocal refractive IOLs: (a) several refractive surfaces with different curvatures. Each one provides a focal point (b) a commercial zonal progressive IOL with five concentric zones on the anterior surface. Each zone repeats the entire refractive sequence corresponding to distance, intermediate and near vision, resulting in wide-range vision.

Hybrid Lenses

Diffractive and refractive optics (Fig. 4) can be combined. In this type of lens, the basic spherical refractive surface is broken up into a refractive zone and a second zone composed of concentric diffractive rings. This combination of zones creates two different focal points for light convergence, one for near objects and one for distant objects. Hybrid IOLs are basically bifocal lenses (Fig. 4). The usual strategy has been to have a distance component with a near component targeting the usual near distance. However, multifocal hybrid lenses are possible.

IOLs AND ACCOMODATION

New research into accommodating intraocular lenses indicates that many patients who get these implants will enjoy good distance and near vision (52). The first hint that intraocular lens implants could accommodate came back in 1986, when Thornton used A-scan biometry to report on anterior movement of a three-piece loop IOL (53). He found that this forward movement allowed some patients to have good uncorrected distance and near vision simultaneously.

The mechanisms of presbyopia remain incompletely understood. A review of the variety of such mechanisms has been presented by Atchison (54). Accommodation in the youthful, phakic human eye is accomplished by contraction of the ciliary body and subsequent release in the resting tension of the zonular fibers by which the crystalline lens is suspended, resulting in increased lens curvature (55–57). The weight of current evidence seems to suggest that although some loss of ciliary body action might contribute to reduced accommodation (58), significant ciliary body function persists into advanced maturity, and that loss of lens and capsule elasticity in concert with changes in the geometry of zonular attachments are probably most culpable in producing the distress of presbyopia (59). If so, replacement of the crystalline lens with a lens that responds to ciliary body contraction should restore accommodative function (60). Attempts to replace the crystalline lens by refilling the capsular bag with appropriately deformable gels have been made (59,61,62). McLeod et al. aimed at designing an accommodating intraocular lens with extended accommodative range that can be adapted to current standard phacoemulsification and endocapsular implantation technique (60). They concluded that a dual optic foldable IOL design can increase the optical effect of a given displacement and suggests improvements for accommodating intraocular lenses.

BIBLIOGRAPHY

1. Barraquer JI. Keratomileusis for the correction of myopia. Arch Soc Amer Oftalmol Optom 1964;5:27–48.

2. Ascher K. Prothetophakia two hundred years ago. Am J Ophthalmol 1965;59:445–446.
3. Nordlohne ME. The Intraocular Lens Development and Results with Special Reference to the Binkhorst Lens, 2nd ed. The Hague: Dr. W. Junk b.v.; 1975: p 14–17
4. Ridley H. Intraocular acrylic lenses—past, present and future. Trans Ophthal Soc UK 1964;84:5–14.
5. Kwitko ML, Kelman CD, editors. Intraocular lens implantation: The Beginnings: The History of Modern Cataract Surgery. The Hague: Kugler Publications; 1998: p 35–52.
6. Apple DJ, Sims J. Harold Ridley and the invention of the intraocular lens. Surv Ophthalmol 1996;40:279–292.
7. Rosen E. History in the making. J Cataract Refract Surg 1997;23:4–5.
8. Choyce P. Intraocular Lenses and Implants. London; H K Lewis & Co. Ltd; 1964.
9. Apple DJ et al. Complications of intraocular lenses: A historical and histopathological review. Surv Ophthalmol 1984;29: 1–54.
10. Javitt JC, et al. Outcomes of cataract extraction with multifocal intraocular lens implantation. Functional status and quality of life. Ophthalmology 1997;104:589–599.
11. Strampeli B. Sopportabilità di lenti acrilichi in camera anteriore nella afachia e nei vizi di refrazione. Ann Oftalmol Clin Ocul 1954;80:75–82.
12. Barraquer J. Anterior chamber plastic lenses. Results and conclusions from five years experience. Trans Ophthalmol Soc UK 1959;79:393–424.
13. Neuhann T, et al. Phakic intraocular lenses, J Refract Surg 1998;14:272–279.
14. Baikoff G, Joly P. Correction chirurgicale de la myopie forte par un implant de chambre anterior dans l'oeil phake. Bull Soc Belge Ophthalmol 1989;233:109–125.
15. Baikoff G. Phakic anterior chamber intraocular lenses. Int Ophthalmol Clin 1991;31:75–86.
16. Rosen ES, Haining WM, Arnott EJ editors. Intraocular Lens Implantation. London, New York: Mosby; 1984.
17. Fyodorov SN, Durnev VV. Operation of dosaged dissection of corneal circular ligament in cases of myopia of mild degree. Ann Ophthalmol 1979;11:1885–1890.
18. Fyodorov SN, Zuev VK, Aznavayez VM. Intraokuliarnaia korrektsia miopii vysokoi stepeni zadnekamernmi otritsatelimi linzami. Oftalmochirugia 1991;3:57–58.
19. Fechner PU, Alpar J. Iris Claw Lens or Lobster Claw Lens of Worst; 1986.
20. Fechner P, Worst J. A New concave intraocular lens for the correction of myopia. Eur J Implant Refract Surg 1989;1:41–43.
21. Hoerauf H, Menz DH, Dresp J, Laqua H. Use of 044 as a solvent for silicone oil adhesions on intraocular lenses. J Cataract Refract Surg 1999;25:1392–1396.
22. Ossipov A. Comparison of internal reflectance patterns of Collamer, acrylic and silicone. 1997. Data on file, STAAR Surgical.
23. Davis EA. Study of post-cataract surgery inflammation with 3 different IOLs (Collamer, SI40NB, AR40). Summary of data found in all patients. Presented at OSN Meeting: New York: October 2003.
24. Jia S, Wang X, Wang E. A study of suitable age for intraocular lens implantation in children according to ocular anatomy and development, Zhonghua Yan Ke Za Zhi. 1996 Sept; 32(5): 336–338.
25. Zou Y, Chen M, Lin Z, Yang W, Li S. Effect of cataract surgery on ocular axial length elongation in young children. Yan Ke Xue Bao 1998;14(1) : 17–20.
26. Flitcroft DI, Knight-Nanan D, Bowell R, Lanigan B, O'Keefe M. Intraocular lenses in children: changes in axial length, corneal curvature, and refraction. Br J Ophthalmol 1999;83(3):265–269.
27. Kora Y, et al. Eye growth after cataract extraction and intraocular lens implantation in children. Ophthalmic Surg 1993;24(7): 467–475.
28. Pan Y, Tang P. Refraction shift after intraocular lens implantation in children. Zhonghua Yan Ke Za Zhi 2001;37(5):328–331.
29. Welford W. Aberrations of Optical Systems. Bristol: Adam Hilger; 1962.
30. Born M, Wolf E. The Diffraction Principles of Optics: Electromagnetic Theory of Propagation, Interference, and Diffraction of Light, 6th ed. New York: Pergamon Press; 1989.
31. Hamam H. Aberrations and their impact on image quality. Wavefront Analysis, Aberrometers & Corneal Topography, Agarwal's edition, 2003.
32. Castejon-Mochon JF, Lopez-Gil N, Benito A, Artal P. Ocular wave-front aberration statistics in a normal young population. Vision Res 2002;42(13):1611–1617.
33. Howland HC, Howland B. A subjective method for the measurement of monochromatic aberrations of the eye. J Opt Soc Am 1977;67(11):1508–1518.
34. Porter J, Guirao A, Cox IG, Williams DR Monochromatic aberrations of the human eye in a large population. J Opt Soc Am A Opt Image Sci Vis 2001;18(8):1793–1803.
35. Paquin MP, Hamam H, Simonet P, Objective measurement of the optical aberrations for myopic eyes, Opt Vis Sci 2002;79: 285–291.
36. Thibos LN, Hong X, Bradley A, Cheng X. Statistical variation of aberration structure and image quality in a normal population of healthy eyes. J Opt Soc Am A 2002;19:2329–2348.
37. Françon M Vision dans un instrument entaché d'aberration sphérique. Thèse, éditions de la Revue d'Optique, 1945.
38. Smith WJ Modern optical engineering, the design of optical system. New York: Me Graw-Hill, 1990.
39. Maréchal A. Étude des effets combinés de la diffraction et des aberrations géométriques sur l'image d'un point lumineux. Revue d'Optique; 1947.
40. Hamam H, New metric for optical performance. Opt Vis Sci 2003;80:175–184.
41. Marsack JD, Thibos LN, Applegate RA Metrics of optical quality derived from wave aberrations predict visual performance. J Vis 2004;4(4):322–328.
42. Tognetto D, et al. Analysis of the optical quality of intraocular lenses. Inv. Ophthalmol & Vis Sci (IOVS) 2004;45/8:2682–2690.
43. Negishi K, Ohnuma K, Hirayama N, Noda T. Effect of chromatic aberration on contrast sensitivity in pseudophakic eyes. Arch Ophthalmol 2001;119:1154–1158.
44. Matin RG. Higher-Order Aberrations and Symptoms with Pseudophakia, Symposium on Cataract, IOL and Refractive Surgery, April 12–16, 2003 San Francisco, CA.
45. Brunette I, et al. Optical quality of the eye with the artisan phakic lens for the correction of high myopia. Optomol Vis Sci 2003 Feb; 80(2):167–174.
46. Liang J, Grimm B, Goelz S, Bille JF. Objective measurement of wave aberrations of the human eye with the use of a Hartmann-Shack wave-front sensor. JOSA A 1994;11:1949–1957.
47. Hamam H. An apparatus for the objective measurement of ocular image quality in clinical conditions. Opt Commun 2000;173:23–36.
48. Hamam H, de Bougrenet JL. Efficiency of programmable quantized diffractive phase elements. Pure and Appl Opt 1996;5:389–403.
49. Wilson K. New Technology Removes Cataract and Improves Vision. Geriatr and Aging 1998;1: p 15.
50. Pieh S, et al. Quantitative performance of bifocal and multifocal intraocular lenses in a model eye: Point spread function in multifocal intraocular lenses. Arch-Ophthalmol. 2002;120(1): 23–28.
51. Malacara D, Malacara Z. Handbook of Lens Design. New York; Marcel Dekker 1994.
52. Karpecki PM. The future of IOLs that accommodate. Rev Opt Dec 2002.

53. Thornton S. Lens implantation with restored accommodation. Curr Cana Ophthal Prac 1986;4:60–62.
54. Atchison DA. Accommodation and presbyopia. Ophthal Physi Opt 1995;15:255–272.
55. Fisher RF. Presbyopia and the changes with age in the human crystalline lens. J Physiol (London) 1973;228:765–779.
56. Koretz JF, Handelman GH. Modeling age-related accommodative loss in the human eye. Math Mod 1986;7:1003–1014.
57. Schachar RA. Zonular function: A new model with clinical implications. Ann Ophthalmol 1994;26:36–38.
58. Hara T, et al. Accommodative intraocular lens with spring action part 1. Design and placement in an excised animal eye. Ophthal Surg 1990;21:128–133.
59. Gilmartin B. The aetiology of presbyopia: A summary of the role of lenticular and extralenticular structures. Ophthal Physiol Opt 1995;15:431–437.
60. McLeod SD, Portney V, Ting A. A dual optic accommodating foldable intraocular lens. British J Ophthal 2003;87:1083–1085.
61. Cumming JS, Slade SG, Chayet A. AT-45 Study Group. Clinical evaluation of the model AT-45 silicone accommodating intraocular lens. Results of feasibility and the initial phase of a Food and Drug Administration clinical trial. Ophthalmology 2001;108:2005–2009.
62. Kuchle M, et al. Implantation of a new accommodating posterior chamber intraocular lens. J Refract Surg 2002;18:208–216.

See also BIOMATERIALS: POLYMERS; CONTACT LENSES; VISUAL PROSTHESES.

LIFE SUPPORT. See CARDIOPULMONARY RESUSCITATION.

LIGAMENT AND TENDON, PROPERTIES OF

G AZANGWE
RM ASPDEN

INTRODUCTION

Tendons and ligaments are fibrous connective tissues that play a mechanical role in the stability and locomotion of the body by transmitting tension. Unlike muscle, which actively contracts, ligaments and tendons are passive. Tendons transmit mechanical forces from muscle to bone, whereas ligaments join bone to bone. Both tendons and ligaments contain relatively few cells (1), and their extracellular matrices are made up of several components. These components are combined in various proportions, and with different organizations to give mechanical properties appropriate to the function of the particular tendon or ligament. There have been a number of reviews in recent years covering specific ligaments, for example, in the rabbit (2), or tendons, such as the human achilles (3), or aspects of their behavior such as healing and repair (4). In this article, the emphasis is on properties the ligaments have in common, which will provide an insight into how and why they behave as they do. This will be based around their functioning as fiber-reinforced materials whose properties are regulated by the cells they contain that produce and maintain the extracellular matrix.

First, this article considers the components of the tissue, not from a biochemical point of view, but as components that may be combined to produce mechanically stable materials. The constituents are considered in terms of the matrix in which are embedded fibers of collagen and varying amounts of elastin. Following this is a discussion of the ways these components interact in ligaments and tendons to yield composite materials with the required mechanical properties. A consideration of some of the ways in which ligaments and tendons may be damaged, and the mechanisms by which they might recover or be repaired, leads to a final, brief review of their surgical replacement. A small section on work being conducted in order to produce a tissue engineered ligament and tendon is also included.

COMPONENTS

Tendons and ligaments are composed primarily of collagen fibers surrounded by a matrix. Here the matrix refers to all the materials that surround the collagen fibers providing both structural support and a medium for diffusion of nutrients and gases. Note this is in contrast to its use in biological terms in which it generally includes the fibrous components. The matrix contains proteoglycans and adhesive glycoproteins and is highly hydrated (typically 65–80% water) (5,6).

Collagen

Collagen fibrils are able to reinforce the weak matrix because of their much greater stiffness and strength in tension (7). The collagen molecule is a long, stiff rod made up of three polypeptide chains wound as a triple helix structure (8). Each fibril is like a rope in which linear molecules are packed together with their axes parallel, within a few degrees, to the fibril axis. Molecules are held together by covalent cross-links so that the fibril is very strong in tension. The regular arrangement of molecules along the axial direction in a fibril gives rise to the characteristic periodicity of 67 nm, which may be seen in electron micrographs (Fig. 1). Although to date, 21 genetically different types of collagen have been identified, types

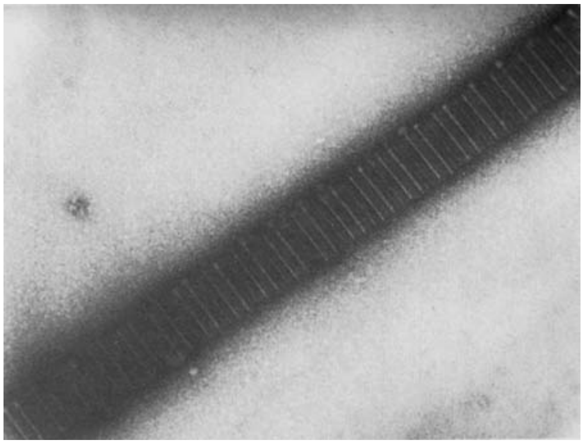


Figure 1. Electron micrograph of collagen fibril from rat tail tendon stained with uranyl formate showing characteristic banding pattern repeated every 67 nm. Micrograph courtesy of Dr D. P. Knight.

I and III fibrous collagens dominate in tendons and ligaments.

There is a heirarchical structure within tendons (9) that has been reviewed often [e.g., (10)], and that appears to have been based on an earlier description of keratin. In this model, collagen molecules are arranged successively into microfibrils, subfibrils, fibrils, and fascicles that are finally packed into the tendon sheath. Evidence for micro- and subfibrils is still equivocal, but a principal mechanical structure is the fibril. These generally have a unimodal distribution of diameters in the newborn, typically a few tens of nanometers, but assume a bimodal distribution in mature tendons and ligaments with mode diameters typically ~100–300 nm (11,12). The ends of fibrils are rarely seen in electron micrographs, and even when they are seen, it is not clear whether they are artifacts of sectioning. Fibrils appear to grow by end-to-end addition of short fibrils and there is evidence from 12-week old mouse skin that fibril tips may eventually fuse with the central region of other fibrils to create a meshwork (13). It is not known whether this happens in tendon or ligament or whether the fibrils are as long as the tendon or ligament. Fibrils are arranged into fascicles, which are ~80–300 μm in diameter and these have a "crimped" morphology that is seen most clearly using polarized light (6,9). This crimp is generally described as a planar zigzag with a sharp change in direction of the collagen fibrils with a periodicity of ~200–300 μm (Fig. 2). On application of a tensile load, initial elongation of the fiber requires a relatively low stress because it simply leads to removal of the crimp (14,15). Once the crimp is removed, the fibers become very stiff. This crimp structure, which is also found in ligaments, explains the characteristic shape of the stress–strain curve for ligaments and tendons (14).

There are many studies of the mechanical properties of tendon and ligament. Most tendons have similar properties, because of their role transmitting forces from muscle to bone and the need to do this most efficiently (16–19). In contrast, every ligament has unique properties that fit it to its function at that particular site in the body. These may range, for example, from the highly extensible ligamentum flavum, helping control spinal posture, to the relatively stiff cruciate ligaments in the knee. Most information on the mechanical properties of collagen has been inferred from experiments on tendon; and though they contain a large proportion of collagen, ~70–80% of the dry weight, the fibrils are surrounded by matrix, and therefore the tissue is really a composite material (6). This makes it difficult to separate the behavior of the individual components.

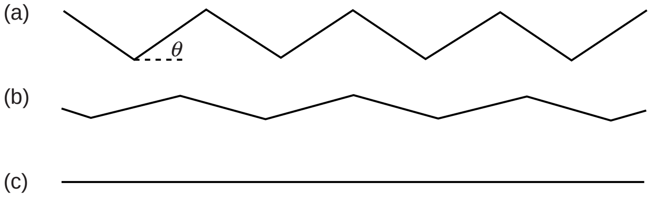


Figure 2. Schematic diagram of the crimp structure in collagen fibers seen from the side. (a) relaxed state [this is grossly exaggerated; the true crimp angle (θ) is ~15°]. As the fiber is stretched, the crimp straightens out (b) until at strains of a ~0.03, the fiber becomes straight (c), at which point it becomes much stiffer.

Proteoglycans

The proteoglycans found predominantly in tendon and ligament belong to the small leucine-rich proteoglycan (SLRP) family; decorin, biglycan, lumican, and fibromodulin though there are small amounts of the large proteoglycans aggrecan (20,21) and versican (21,22). The SLRPs all comprise a repeating structure that is rich in leucine residues, between 13 and 16% of all residues (23). They are present in different tissues in differing amounts and appear at different stages of development (24). Their function is poorly understood though gene knockout studies in mice have shown marked osteopenic effects on bone, skin laxity, and irregular collagen fibers in tendon (25). Most of these SLRPs have one or two glycosaminoglycan (GAG) chains of varying lengths attached, generally either dermatan sulfate or chondroitin sulfate, which can interact electrostatically with collagen (26,27). The proteoglycan decorin has been found to be localized in the tissue to a specific region of a collagen fibril (28). It is also reported to play a regulatory role in collagen fibrillogenesis, by affecting fibril radius (13), and increases the strength of uncrosslinked fibers (29,30). In regions where tendons pass over bone and are subjected to compressive loading in addition to tension, a fibrocartilaginous tissue containing large amounts of aggrecan and biglycan develops (31,32). The adhesive glycoproteins include fibronectin and thrombospondin (33–35), both of which contain possible attachment sites for cells. In humans, these were reported to be more common in the tendon sheath than in the fibrous bulk (36) and fibronectin was found to be up-regulated at sites of injury (34).

The matrix is weak in shear; that is, if it is loaded in compression, it tries to slide out sideways unless it is contained. This behavior may be readily seen in jelly (Jello in the United States) and is not surprising given its high water content, since fluids simply flow when sheared. It is not easy to measure the shear strength of matrix. Proteoglycans at similar concentrations have a very low shear strength (37); however, matrix may be stiffer than this because of the interactions between its macromolecular components. An analysis of the behavior of tendon suggests that its matrix would have a shear modulus of ~100 kPa (38). Because of this low stiffness in shear, the matrix alone is not well suited to bearing loads. Also, its proportion in ligament and tendon is quite low, ~10–20% of the dry weight. The ways in which matrix may transmit stress to the fibers and prevent crack propagation will be discussed later.

Elastic Fibers

Electron microscopy reveals the presence of elastic fibers in ligaments and tendons (39,40). Elastic fibers have two components; elastic fiber microfibrils and elastin. The microfibrils have a diameter of 10–12 nm and consist of glycoproteins. Elastin is composed of mainly hydrophobic nonpolar amino acids with a high content of valine (41).

Elastic fibers are highly extensible, they do not creep and their extension is reversible at high strains. Their mechanical properties are thus very different from collagen. Most of our knowledge of elastic fibers comes from experiments on ligamentum nuchae, a ligament from the cervical spine, which contains ~70% elastin by dry weight (42). Elastin closely resembles a rubber in many respects and its mechanical properties are certainly very similar (43). Purified samples of ligamentum nuchae will extend to roughly twice their resting length before breaking.

The extensibility of a tendon or ligament depend in part on the elastin content of the tissue. Ligamentum flavum, from the spine, which may typically be strained to ~50% contains roughly 70% elastin, by dry weight (44), whereas tendon, which works at strains <4% contains only 2% elastic fibers by dry weight (1). It is fairly easy to see why highly extensible tissues have a high proportion of elastin, but not quite as easy to explain the presence of elastic fibers in a relatively inextensible tissue such as tendon. A clue is provided by some synthetic fibrous composite materials that contain two different kinds of fiber (45). Here a small proportion of strong, low stiffness fibers added to the composite produces a material that is less susceptible to failure under sudden application of load than one that contains only stiff fibers; that is, it makes the material less brittle. It may be that the small proportion of elastic fibers in tendon provide some protection against the sudden application of load that may occur, for example, if an animal is startled.

Fiber–Matrix Interactions

The combination of strong fibers in a weak matrix leads to materials that are less susceptible to mechanical damage while maintaining a high proportion of the strength of the fibers. In particular, they are less susceptible to sudden failure than a homogeneous material would be; a property called “toughness” (46). This composite nature has been recognized for many years (6) and provides a theoretical framework for understanding the properties of the tissues. It also enables some useful comparisons to be made between relatively simple synthetic composites and biological tissues in which the complexities of composition and structure make modeling very difficult. The aim is to obtain an understanding of how the similarities in the tissues, fibers in a matrix, enable them to function in general terms before considering the differences (in composition, and organization), which give them their specific properties. The function of collagen fibrils and fibers in such a composite is to withstand axial tension, since, like any rope, they have little resistance compression and flexion (7). As the tissue is stretched the matrix will try to flow and this will exert a shear force along the surface of the collagen fibers tending to stretch and orient them (7,47). This length increase, which is normally expressed as a fraction of the original length and is then termed “strain”, leads to a restoring force in the fiber that balances the applied force. The behavior is rather like a loaded spring that stretches to enable it to bear load, but returns to its relaxed length on removing the load. Similarly, collagen fibers are able to reinforce a tissue if they are oriented so that an applied load tends to stretch them. The nature of the shear force exerted by the gel is unknown, but two simple models, those of elastic and plastic (or frictional) stress transfer, have been used to investigate stresses in the fibers and the force that has to be generated at the fiber surface to enable them to function in this way (47–49). Fibers that are shorter than the tissue are still able to provide reinforcement (50). Some fibrils observed in tissues (51) and those grown *in vitro* (52,53) appear to be tapered, rather than having a uniform radius. Analytical and finite element models of idealized single-fiber composites have shown that tapered fibers have two distinct advantages over uniform fibers: the axial stresses within the fiber are more uniformly distributed and they contain a much smaller amount of material, though their effectiveness at reinforcing is just as great. A more uniform stress within the fiber means more of the fiber is carrying a significant stress, thus making better use of the fiber, and avoids the generation of stress concentrations that are potentially damaging and could lead to fiber fracture. In addition, the volume of material in a cone, for example, is only one-third of that in a straight cylinder and, therefore, a tapered fiber incurs a far smaller metabolic cost by the cells to produce it. In straight-cylindrical fibers it has been calculated that interactions at the fiber surface do not have to be great in order to load fully the fiber. In tendon, assuming conservatively only one interaction per 67 nm D-period, it was estimated that fiber–matrix interaction forces of the order of only 10 pN was sufficient to load fully the fiber (47). These forces are similar in magnitude to van der Waals forces or hydrogen bonds. This suggests that permanent bonds or covalent interactions between fiber and matrix are not essential for the mechanical functioning of the tissue though, of course, it does not preclude them. Regulating the interaction between fibers and matrix is clearly important in this model of how the tissues function and decorin, as described above, is a prime candidate for a role in this. Changes in the concentration and orientation of collagen and its interactions with the matrix have been used to explain the dramatic changes in a similar fibrous tissue, the uterine cervix, that occur during parturition (54). The presence of the matrix around the collagen fibrils is also important when it comes to preventing crack propagation. This will be considered in more detail in the context of the tissues themselves.

LIGAMENT

Ligaments are short bands of tough, but flexible, fibrous connective tissue that bind bones together and guide joint motion, or support organs in place. The word ligament is derived from the Latin word “ligare,” which means to bind. Generally, ligaments can be classified into two major subgroups. There are those connecting the elements of the skeletal system (usually crossing joints) and those connecting other soft tissues, such as the suspensory ligaments in the abdomen. This section only considers skeletal ligaments. The main function of the skeletal ligaments, such as the anterior cruciate ligament (ACL) of the knee joint, is to stabilize and control normal kinematics, to prevent

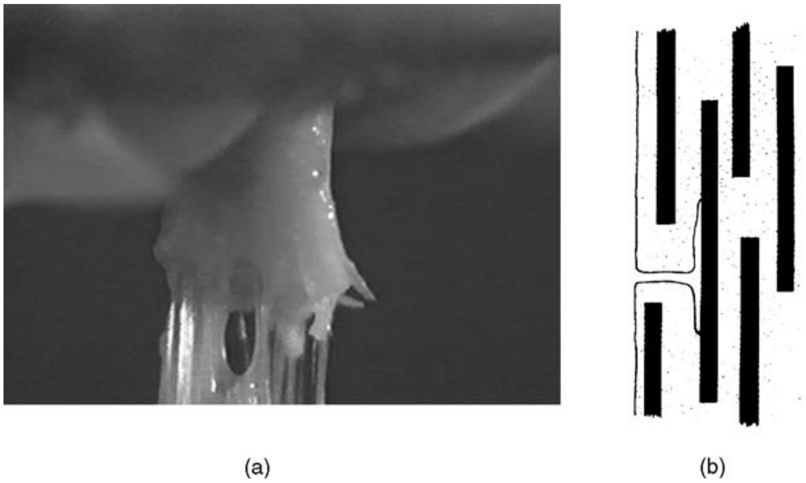


Figure 3. (a) Video image showing a ligament failing under tensile strain. (b) A schematic diagram showing a longitudinal section through a fiber composite illustrating how a weak matrix prevents crack propagation from one fiber to another and dissipates its energy by creating cracks in directions other than across the composite.

abnormal displacements and rotation that may damage the articular surfaces (2,55). At the insertion to bone, the ligaments change from flexible ligamentous tissue to rigid bone, mediated by a transitional zone of fibrocartilage and mineralized cartilage. This helps to prevent stress concentration at the attachment site by allowing gradual change in stiffness (56,57).

Similar to tendon, ligaments are primarily composed of collagen embedded in a weak matrix (7). The collagen molecules in ligaments pack together to form fibrils and the fibrils aggregate into larger fibrous bundles (2). As described previously, the function of collagen fibrils is to provide tensile reinforcing for the matrix. The proportion of fibers within the ligamentous structure and the orientation of the collagen fibers are the main factors that govern the mechanical behavior of the tissue (7). In contrast to tendon, there is commonly less collagen, which is less highly oriented than in tendon, conferring a generally greater extensibility to these tissues.

The collagen fibrils also prevent damaged tissues from failing suddenly. For example, most ligaments do not tear straight across when they are damaged (Fig. 3). Instead, small tears in the matrix are diverted when they encounter the strong collagen fibrils (see below on failure). There is then a possibility that a damaged ligament can heal while, in the meantime, retaining the ability to withstand some load. Some ligaments, however, (e.g., the ACL), have a limited ability to heal when ruptured and this means that it often needs to be replaced or reconstructed when ruptured. A brief summary of different options available for treating ruptured ligaments will be presented in a later section.

Unlike tendons that all have very similar composition, structure, and function, the same cannot be said of ligaments, and it is far harder to make general comments about their properties. Much less is known, too, about the relationship between their structures and functions as the arrangements of their collagen fibrils are more complex than those in tendons (58,59). This greater complexity is understandable when it is realized that the function of a ligament is very dependent on its position in the body; for example, the medial collateral ligament in the knee of a sheep operates at strains of ~0.02 (60), whereas the ligamentum flavum of the human spine operates at strains of up to 0.6. Figure 4 shows that ligamentum flavum is much less stiff than tendon.

It is not surprising that some ligaments contain a high proportion of elastin (~60–70% of the dry weight), which enables them to withstand the high strains to which they are subjected without fracture (61). Ligaments are viscoelastic, that is their properties are time dependent and they appear stiffer if stretched more rapidly. These ligaments exhibit hysteresis, that is, they lose energy on being taken through a cycle of stretching and relaxing. Tkaczuk (61) published a detailed account of the mechanical properties of longitudinal ligaments from the human spine. These deform elastically up to strains of ~0.25, when the stress is ~5 MPa, and rupture at a stress of ~20 MPa. Shah et al.(62) also showed that, like tendons, the collagen fibers are crimped and this crimp disappears at strains of ~1.2–2.8% depending on the ligament. When ligaments are cut from the joint, they can often be seen to contract rapidly, suggesting that they are held in a state of tension even when the joint is in a relaxed state.

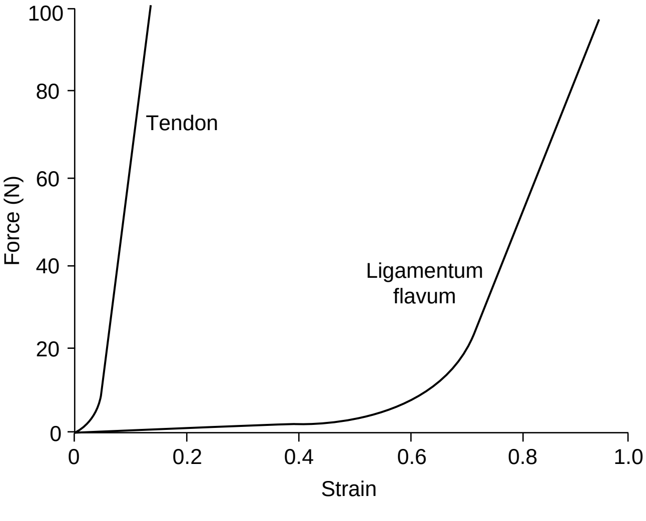


Figure 4. Comparison of force–strain curves obtained for extension of tendon and ligamentum flavum.

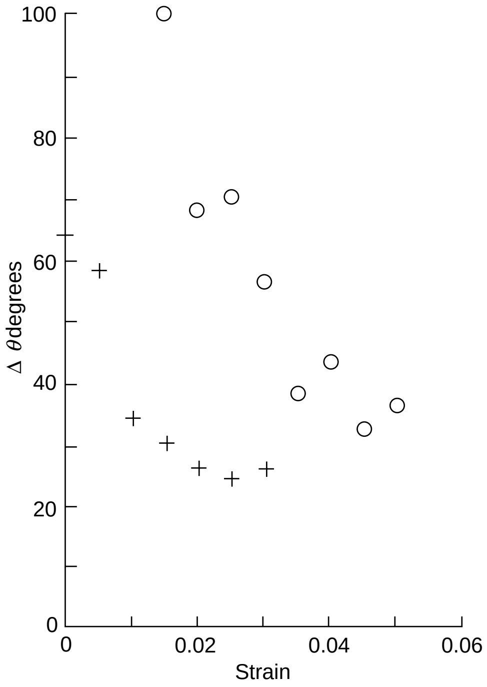


Figure 5. Full width at half-maximum (fwhm), Δ, of the distribution of orientations of collagen fibers in ligamentum flavum. (o) and posterior longitudinal ligament, (+) as a function of strain.

Spinal ligaments provide a good example of the mechanical function of ligaments in a joint (58). The longitudinal ligaments and the ligamentum flavum act together with the intervertebral disc to achieve a mechanically stable joint and serve to limit its mobility. The ligamentum flavum is almost twice as far from the axis of rotation in forward bending as the posterior longitudinal ligament, and hence it can be seen that it needs to be roughly twice as extensible (58,63). This is partly explained by a higher elastin and lower collagen content in ligamentum flavum (61), but also by a less highly aligned organization of collagen fibers (63). As the ligament is stretched, the fibers become more highly aligned and this will increase the stiffness of the tissue, that is its resistance to extension. X-ray diffraction experiments have measured the spread of orientations of fibrils in these ligaments and modelling has shown how this decreases with increasing extension (47). Figure 5 shows that the width of the distribution, Δθ, as measured at half the peak height, is greater for ligamentum flavum than posterior longitudinal ligaments. This mechanism provides an explanation for the form of the force–strain curve, shown in Fig. 4, One final point about ligaments is that they have a nerve supply that makes them potential sources of pain. It has also been suggested that they may function as proprioceptors as part of a reflex arc, that is, the ligaments would act as sensors to detect the position of a joint and the information would then be used to control the muscles around the joint thereby controlling its movement and stability (64).

In summary, ligaments are composite materials containing crimped collagen fibers that are prestressed in the relaxed joint. They have a nonlinear stress–strain curve and are viscoelastic. The collagen fibers are relatively disoriented in the unstretched tissue and become more highly aligned as the tissue is stretched. They often contain a proportion of elastin. Their composition and structure depend on their position in the body and their dynamic behavior, that is, the change in structure with strain, becomes more important.

TENDON

The function of tendon is to transmit the force generated by a contracting muscle to the correct point of application on a bone so as to manipulate a joint. Tendons are often preferable to direct attachment of muscle to bone because of various functional requirements. Muscles have a low tensile strength, defined as load at fracture per unit cross-sectional area. This means that they must have a large cross-sectional area in order to transmit sufficient force without tearing. Around many joints (e.g., the fingers), there is insufficient space to attach many muscles. The muscle, therefore, is located further away and attachment made by a tendon, which may be tens of centimeters long in the hand and forearm. Tendons, therefore, need to be strong, so that they can be relatively slender, and stiff, so that the force developed by the muscle is transmitted to the bone without energy being wasted on stretching the tendon. A graph of force as a function of strain for a tendon is shown in Fig. 6. This type of curve is obtained by subjecting the tendon to various forces and recording the amount by which the tendon is stretched, and many of the early studies have never been bettered (6,9,18,40,65). The steepness of the stress–strain curve is a measure of the stiffness of the material. Figure 6 shows that for small strains the tendon requires very little force to stretch it but thereafter it stiffens considerably. The stress at this point is ~10 MPa, which is several orders of magnitude greater than the stresses needed to shear the matrix (6). If the stiffness did not increase, a muscle would continue to stretch the tendon and the force would not be transmitted to the bone.

Tendons are believed to function in the body at strains up to ~0.04 (66,67). Beyond this strain, a tendon does not return to its original length when the applied stress is removed. A tendon will break at strains of ~0.1 (67). The

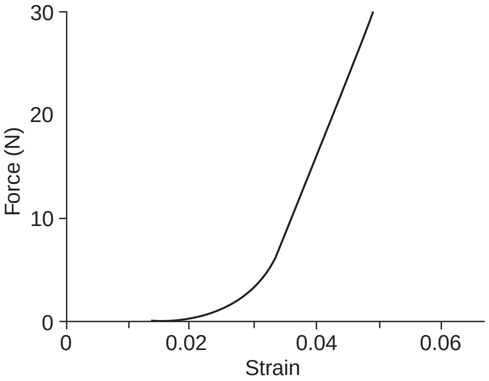


Figure 6. Force–strain curve for tendon.

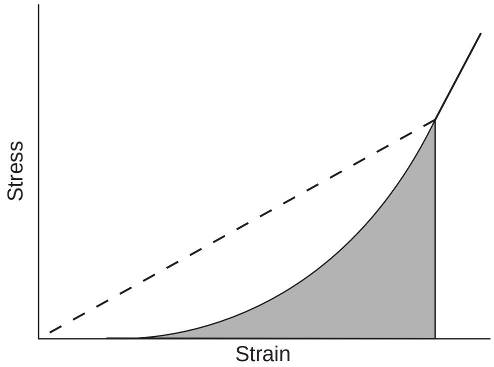


Figure 7. Schematic of stress–strain showing energy stored in a stretched tendon.

initial stages of tendon extension involve straightening the crimp of the collagen fibers described previously.

The energy stored in the stretched tendon is given by the area under the stress–strain curve, as shown in Fig. 7, and it is this energy that must be used to create a tear in the tissue. This energy is lower in a material with a J-shaped stress–strain relationship than if, for example, it were linear. Minimizing the energy available to cause a fracture in this way gives the tissue a property known as resilience, that is, a tendon does not suddenly fail if it is overloaded—unlike a steel wire. While the crimped collagen fibers are being straightened, the weak matrix must be sheared. Because of the fluid-like nature of the matrix, it tends to flow. The rate of flow depends on the force applied to it and the amount of flow is greater the longer the force is applied. The result is that tendons are "viscoelastic" (18,38,68). The effect of the rate at which force is applied to the tendon is shown in Fig. 8.

This time dependence of mechanical behavior leads to a phenomenon known as "creep". For example, when a load of 10 N was applied to a human flexor digitorum tendon, the initial strain was 0.015, but 100 s later under the same load, the strain had increased to 0.016 (69). Thus, if a tendon is stretched rapidly, the matrix has less chance to flow and creep in the material is less resulting in the tendon being stiffer as shown in Fig. 8. Viscous flow of the matrix also provides a mechanism for dissipating energy; more work is done in stretching a tendon than is recovered when the tendon is allowed to relax. This phenomenon is known as hysteresis and is illustrated in Fig. 9. The behavior of a tendon is therefore intermediate between that of a steel wire that stores all the energy used to stretch it, and a viscous liquid that simply flows to a new position and does not store any of the energy put in to cause it to flow.

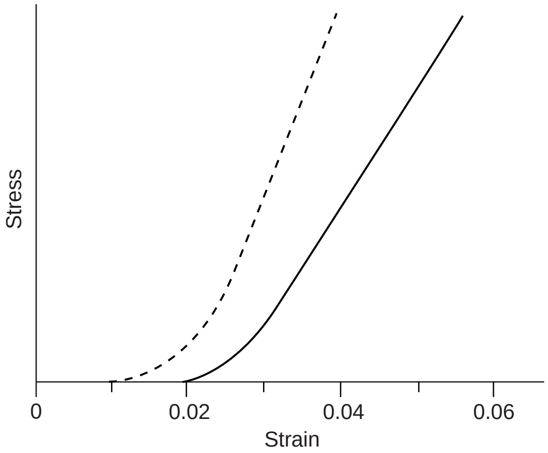


Figure 8. Schematic stress–strain curves for tendon to show the effects of different loading rates. These curves correspond to slow loading (continuous curve) and rapid loading (dashed curve).

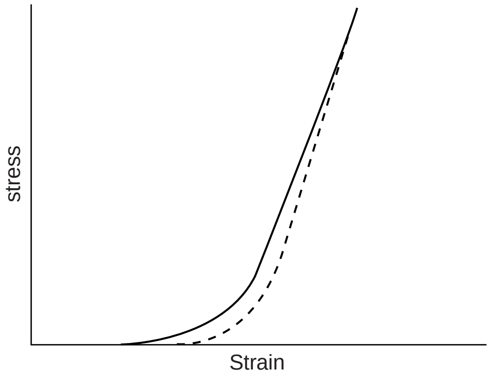


Figure 9. Stress–strain curves for tendon that is stretched (continuous curve) and then relaxed (dashed curve), showing hysteresis; that is, not as much energy is recovered on relaxing as was initially used to stretch the tissue.

MEASURING THE PROPERTIES OF LIGAMENTS AND TENDONS

To quantify the physical properties of ligaments and tendons, mechanical testing of bone–ligament/tendon–bone complexes is often performed (70–73). That this method is often used is partly due to the difficulty in testing isolated ligaments and tendons. Ideally, testing isolated ligaments and tendons would provide measures of the material properties of the tissue alone, but such tests are complicated by difficulties in effectively securing the cut ends (2). Putting the free ends in clamps often results in stress concentrations at the grips, which may contribute to premature failure. Although the use of bone–ligament/tendon–bone complexes still provides more secure clamping, it also increases the difficulty of separating the properties of the ligament from those of the insertion sites. When such complexes are subjected to tensile loading, the resulting load-displacement curve represents the mechanical properties of the bone–ligament/tendon–bone complex as whole rather than specifically about the material that makes up the ligament or tendon.

In order to obtain the material properties of the ligaments, one needs to measure their length (to calculate strain from deformation divided by original length) and cross-sectional area (to calculate stress from applied force divided by original area). From stress–strain curves material properties such as Young's modulus (slope of stress–strain curve), maximum stress, maximum strain, and energy density (area under stress–strain curve) can be determined. Special devices such as buckle transducers (74) and Hall-effect displacement transducers (75) have been used to measure ligament strains during testing. The drawback of such devices is that they rely on direct contact with the tissue sample and that may influence the results. Optical analysers have also been employed as a noncontact method to measure ligament strains (2). However, inaccuracies may occur because the irregular dye blobs used as markers change shape on stretching making it difficult to define unique points. Another technique that has been employed to measure strain is the use of video dimension analyser (VDA) (76,77). This method requires no direct contact with the specimen, but relies on a recorded video image.

The irregular and complex shape and geometry of ligaments and tendons also make it difficult to measure their cross-sectional area. Although flexible callipers, which are able to follow contours better, have been used to measure ligament cross-sectional areas (2), they still require contact, and this results in errors in measurements. Other investigators have calculated the cross-sectional area of a known length of ligament from measurements of its density by floatation in a mixture of xylene and carbon tetrachloride (78). A number of noncontact methods, such as the use of a rotating microscope (79), use of the VDA (80) and the laser micrometer (81) have also been employed to measure the cross-sectional area.

When mechanical properties of ligaments and tendons are being determined, it is important to consider the rate at which they are loaded since they are viscoelastic, that is, their mechanical properties depend on the rate at which they are deformed (82–86). This sensitivity to strain rate means that ligaments and tendons exhibit properties of stress relaxation (decreased stress with time under constant deformation) and creep (increased deformation with time under constant load) (87,88).

Because of the complex geometry of tendons and ligaments, the orientation of the specimens during mechanical testing affects their physical properties and the manner in which they fail, and should therefore be taken into account when performing mechanical tests. Torsion has been implicated as a factor in the rupture of the ACL during sporting injuries (89–91). Cyclic loading has also been found to lower the yield point or soften the ligament by increasing its compliance (decrease the slope of the linear region of the stress–strain curve) (55). Azangwe et al. (92) showed that, when combined, tension–torsion loading affects both structural and mechanical properties of anterior cruciate ligaments.

LIGAMENT AND TENDON FAILURE MECHANISMS

Since ligaments and tendons consist of collagen fibers reinforcing a weak matrix, it is reasonable to compare their behavior under tensile loading with that of synthetic fiber-reinforced composites, since their failure mechanisms are well established. This section describes some of the modes of failure of fiber-reinforced composites and how they are related to those of the ligaments and tendons. Detailed accounts of failure modes of synthetic fiber-reinforced composites can be found in textbooks, for example, Agarwal and Broutman (45), and Kelly and Macmillan (50). When a material is subjected to any kind of loading, it can absorb energy by two basic mechanisms:

1. Material deformation.
2. Creation of new surfaces.

Material deformation occurs whenever a material is subjected to load. However, if the energy supplied is sufficiently large, cracks may be initiated. Whether they propagate depends on the relative amounts of energy required to create new surfaces compared with that stored in the deformed matrix. For brittle materials such as glass, the energy required to create a new fracture surface is small and though only a small amount of deformation takes place, this is elastic and the associated energy is sufficient to propagate the crack. This means that brittle materials have a low energy-absorption capability. On the other hand, for ductile materials, large plastic deformations occur, which dissipate energy, resulting in large energies being absorbed rather than being available to drive a fracture. This finding shows that the total energy-absorbing capability or "toughness" of a material can be enhanced by increasing either the length of the path of the crack during separation or the material-deformation capability. In metals, the latter mechanism frequently occurs and metallurgical processes are developed to maintain ductility. In composites, replacing low energy-absorbing constituents with greater energy-absorbing constituents can enhance the toughness.

As is the case with many materials, failure in a fiber-reinforced composite emanates from small inherent defects in the material. Several failure events may occur during the fracture of a fiber reinforced composite material, such as,

1. Microcracking of the matrix.
2. Separation of fibers from the matrix (debonding and pull-out).
3. Breaking of fibers,

Forcing a crack to take a longer path is the main mechanism encountered in fiber composites to increase toughness. In fibrous materials, when a matrix crack encounters a strong fiber placed perpendicular to the direction of crack propagation, if the crack cannot cross the fiber because it is too strong, the crack is forced to branch to run parallel with the fiber (Fig. 10). If the crack goes right around the fiber, this may result in fiber debonding, and pull-out, that is, becoming detached from the matrix and pulled out leaving fiber ends showing as the material is stretched. In many cases the surface area produced by secondary cracks is much larger than the area of the primary cracks. This may increase the fracture energy many times and is an effective way of increasing the toughness of composites or the total energy absorbed during fracture. Fibers may eventually fracture when their strength is exceeded. For most synthetic fiber-reinforced composites, fibers are separated by matrix and therefore are unable to pass energy directly

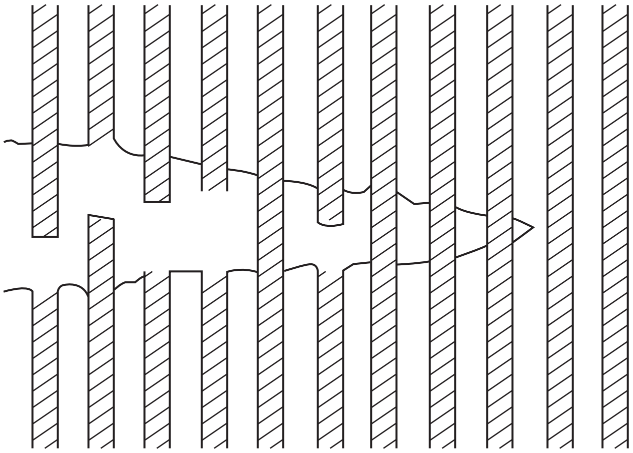


Figure 10. Model of crack tip in a fiber-reinforced composite showing how a crack may propagate Fibers become debonded from the matrix and are pulled out as the crack widens.

from one to another, hence it is unlikely that they all fail together. Collagen fibers in ligaments are similarly separated by the matrix, and hence the transmission of stress from fiber to fiber in the ligament is indirect (7). The situation is more complicated in biological tissues because of the complexity of the components and their arrangement within the tissue and because of the possibility of repair. It seems reasonable to assume that a crack will start first in the weak matrix rather than in the strong collagen fibers. It is unlikely that the crack will spread into a fibril, as previously discussed, but it will be deflected by the fibrils into new directions.

It appears that tendons and ligaments can continue to withstand stress long after they are damaged, but before complete fracture occurs (70,71,93). However, since damage implies an irreversible change, at least until the biological repair process begins, the tissue will not return to its original dimension when the stress is removed. When testing bone–tendon/ligament–bone complexes, an additional failure mode may occur at the ligament insertion to bone. For example, the ligamentum flavum tears at the enthesis, the junction with the bone, leaving virtually no possibility of natural healing (72). Fortunately, this injury does not appear to occur in vivo and can occur in isolation or in combination with other failure modes. As mentioned previously, most ligaments appear to be prestressed within the body so that if they are severed or avulsed they retract and the damaged ends are no longer in contact. This makes it very difficult for cells within the tissue then to bridge the gap by synthesizing new matrix.

The tearing of tendons is a fairly common injury, though rupture of the tendon tends to occur at the junction with bone. Healing of torn tendons in humans is slow, and is not always improved by surgical intervention (3). Healing starts with the invasion of cells into the damaged area that first lay down fine, poorly oriented collagen fibrils (94). These fibrils later increase in diameter and become increasingly oriented as the cell population drops and the structure becomes more akin to that of the original tendon (95).

Damage to and repair of ligaments follows a very similar pattern to that described for tendons. A point to note in all this is that there are differences in healing characteristics between different ligaments. The ACL of the knee joint, for example, appears to have a poor healing capacity when injured prompting the need for reconstruction. At the junctions of ligaments to bone, the fibers of the ligament become more compact, then cartilaginous and finally calcified before finally merging into the bone and there are changes in cell phenotype and expressed proteins (96,97). This complex structure reflects the difficulty of attaching a tough, flexible material to a hard, brittle one.

LIGAMENT AND TENDON REPAIR

Because of the slow rate of healing of certain ligaments and tendons, their reconstruction with synthetic materials has been attempted, with varying degrees of success, for many years (98). The area of prosthetic materials and methods is so vast that only a very brief survey will be attempted here. The main approaches that have been tried are replacement using tissues taken from another part of the body, or from an animal, and complete substitution by a synthetic material. None of these approaches has so far proved entirely satisfactory. Success rates of 80–90% are reported for both techniques (99–101), but some long-term studies have shown that this success rate may fall to 40–50% after ~5 years (102) and there are reports of high, ~50%, incidence of degenerative change (101,103). Friedman et al. (104), reviewed the autogenous reconstruction of the ACL using patellar tendon, iliotibial band, gracilis tendon, semitendinosus, or meniscus that gives some idea of the range of tissues that have been tried. Unfortunately, most repairs stretch with time resulting in loss of stability (105); this is probably due to the differences in structure and mechanical properties between the original tissue and the replacement, described earlier. If the replacement tissue is stretched too much while being fixed in position, it may be irreversibly strained.

Early attempts at the synthetic replacement of ligaments and tendons, using silk, for example, were not successful. These prostheses were intended to be permanent, but had not the strength and fatigue resistance to withstand the millions of cycle of loading imposed on them during the lifetime of the recipient. More recently polyester (106), carbon fiber (107,108), or various combinations of synthetic materials and autogenous tissue have all been tried but still seem not to overcome this particular problem (103,109,110).

Another approach that is currently being explored for augmenting or reproducing ligaments and tendons is that of tissue engineering (111,113), though this has not yet reached the stage of clinical utility. These techniques may include the development of biodegradable scaffolds, on which it is hoped to encourage cells from the patient to grow a replacement tissue (113), and growth factors (114) and their introduction into the tissue using gene transfer (112,115).

SUMMARY

Tendons and ligament are connective tissues subject primarily to tensile forces. They comprise crimped collagen fibrils and some elastin embedded in a weak matrix. Collagen fibers in tendon, composed of bundles of fibrils, are highly aligned along the direction of applied force. The initial structural response to the application of force is straightening of the crimp, which occurs at a strain of ~2%, after which the tendon stiffens considerably. Higher strains are not immediately reversibly and may lead to structural damage to the tissue. The function of the crimp appears to be to minimize the energy stored in the stretched tissue thus reducing the energy available to cause fracture. Many ligaments can be stretched more than tendons, because of their more complex collagen fibril organization and, sometimes, the high proportion of elastin present. Tendons and ligaments are viscoelastic; they do not store all the energy used to stretch them, their response to load depends on the rate at which it is applied, and they continue to deform even if the applied load remains constant. Because a lot of energy is required to produce a large fracture surface, they do not break easily, that is, they are tough materials. Because some ligaments do not heal well when injured, there is a need to replace them. However, success in this area is still limited. Future replacements may include tissue engineered ligaments.

ACKNOWLEDGMENTS

We thank Professor D.W.L. Hukins for valuable discussions over many years and the many other colleagues who have contributed to some of the work we have described here. We also thank the Medical Research Council, the Engineering and Physical Sciences Research Council, and the Arthritis Research Campaign for financial support.

BIBLIOGRAPHY

1. Fank CB, Shrive NG. Ligament. In: Nigg BM, Herzog W, editors. Biomechanics of the Musculo-Skeletal System. Chichester: J Wiley; 1999.
2. Frank CB, Hart DA, Shrive NG. Molecular biology and biomechanics of normal and healing ligaments—a review. Osteoarthritis Cartilage 1999;7:130–140.
3. Maffulli N. Rupture of the Achilles tendon. J Bone Joint Surg 1999;81-A:1019–1036.
4. Woo SL, Debski RE, Zeminski J, Abramowitch SD, Saw SS, Fenwick JA. Injury and repair of ligaments and tendons. Annu Rev Biomed Eng 2000;2:83–118.
5. Nachemson AL, Evans JH. Some mechanical properties of the third human lumbar interlaminar ligament (ligamentum flavum). J Biomech 1968;1:211–220.
6. Elliott DH. Structure and function of mammalian tendon. Biol Rev 1965;40:392–421.
7. Hukins DWL, Aspden RM. Composition and properties of connective tissues. Trends Biochem Sci 1985;10:260–264.
8. Brodsky B, Ramshaw JA. The collagen triple-helix structure (Review). Matrix Biol 1997;15:545–554.
9. Kastelic J, Galeski A, Baer E. The multicomposite structure of tendon. Connect Tissue Res 1978;6:11–23.
10. Vincent JFV. Structural Biomaterials Princeton. 2nd ed. New Jersey: Princeton University Press; 1990.
11. Parry DA, Barnes GR, Craig AS. A comparison of the size distribution of collagen fibrils in connective tissues as a function of age and a possible relation between fibril size distribution and mechanical properties. Proc R Soc London Ser B Biol Sci 1978;203:305–321.
12. Patterson-Kane JC, Wilson AM, Firth EC, Parry DA, Goodship AE. Comparison of collagen fibril populations in the superficial digital flexor tendons of exercised and nonexercised thoroughbreds. Equine Vet J 1997;29:121–125.
13. Kadler KE, Holmes DF, Graham H, Starborg T. Tip-mediated fusion involving unipolar collagen fibrils accounts for rapid fibril elongation, the occurrence of fibrillar branched networks in skin and the paucity of collagen fibril ends in vertebrates. Matrix Biol 2000;19:359–365.
14. Buckley CP, Lloyd DW, Konopasek M. On the Deformation of Slender Filaments with Planar Crimp: Theory, Numerical Solution and Applications to Tendon Collagen and Textile Materials. Proc R Soc London Ser A, Math Phys Sci 1980; 372:33–64.
15. Diamant J, Keller A, Baer E, Litt M, Arridge RGC. Ultrastructure of collagen as a function of ageing. Proc R Soc London Ser B Biol Sci 1972;180:293–312.
16. Betsch DF, Baer E. Structure and mechanical properties of rat tail tendon. Biorheology 1980;17:83–94.
17. Haut RC, Lancaster RL, Decamp C. Mechanical properties of the canine patellar tendon—some correlations with age and the content of collagen. J Biomech 1992;25:163.
18. Woo SL. Mechanical properties of tendons and ligaments. I. Quasi-static and nonlinear viscoelastic properties. Biorheology 1982;19:385–396.
19. Woo SLY, Sites TJ. Current advances on the study of the biomechanical properties of tendons and ligaments. In: Nimni ME, editor. Collagen, Volume II, Biochemistry Biomechanics. Boca Raton: CRC Press; 1988.
20. Campbell MA, Tester AM, Handley CJ, Checkley GJ, Chow GL, Cant AE, Winter AD, Cain WE. Characterization of a large chondroitin sulfate proteoglycan present in bovine collateral ligament. Arch Biochem Biophys 1996;329:181–190.
21. Robbins JR, Vogel K. Regional expression of mRNA for proteoglycans and collagen in tendon. Eur J Cell Biol 1994;64: 264–270.
22. Thomopoulos S, Hattersley G, Rosen V, Mertens M, Galatz L, Williams GR, Soslowsky LJ. The localized expression of extracellular matrix components in healing tendon insertion sites: an in situ hybridization study. J Orthop Res 2002;20:454–463.
23. Hocking AM, Shinomura T, McQuillan DJ. Leucine-rich repeat glycoproteins of the extracellular matrix. Matrix Biol 1998;17:1–19.
24. Wilda M, Bachner D, Just W, Geerkens C, Kraus P, Vogel W, Hameister H. A comparison of the expression pattern of five genes of the family of small leucine-rich proteoglycans during mouse development. J Bone Miner Res 2000;15:2187–2196.
25. Corsi A, Xu T, Chen XD, Boyde A, Liang J, Mankani M, Sommer B, Iozzo RV, Eichstetter I, Robey PG, Bianco P, Young MF. Phenotypic effects of biglycan deficiency are linked to collagen fibril abnormalities, are synergized by decorin deficiency, and mimic Ehlers-Danlos-like changes in bone and other connective tissues. J Bone Miner Res 2002;17:1180–1189.
26. Gelman RA, Blackwell J. Collagen-mucopolysaccharide interactions at acid pH. Biochim Biophys Acta 1974;342: 254–261.
27. Lindahl U, Hook M. Glycosaminoglycans and their binding to biological macromolecules. Annu Rev Biochem 1978;47: 385–417.
28. Scott JE, Orford CR. Dermatan sulphate-rich proteoglycan associates with rat tail tendon collagen at the d band in the gap region. Biochem J 1981;197:213–216.
29. Weber IT, Harrison RW, Iozzo RV. Model structure of decorin and implications for collagen fibrillogenesis. J Biol Chem 1996;271:31767–31770.

30. Pins GD, Christiansen DL, Patel R, Silver FH. Self-assembly of collagen fibers. Influence of fibrillar alignment and decorin on mechanical properties. Biophys J 1997;73:2164–2172.
31. Evanko SP, Vogel KG. Proteoglycan synthesis in fetal tendon is differentially regulated by cyclic compression *in vitro*. Arch Biochem Biophys 1993;307:153–164.
32. Koob TJ, Clark PE, Hernez DJ, Thurmond FA, Vogel KG. Compression loading *in vitro* regulates proteoglycan synthesis by tendon fibrocartilage. Arch Biochem Biophys 1992; 298:303–312.
33. Cockburn CG, Barnes MJ. Characterization of thrombospondin binding to collagen (type I) fibers: role of collagen telopeptides. Matrix 1991;11:168–176.
34. Amiel D, Gelberman R, Harwood F, Siegel D. Fibronectin in healing flexor tendons subjected to immobilization or early controlled passive motion. Matrix 1991;11:184–189.
35. Kannus P, Jozsa L, Jarvinen TA, Jarvinen TL, Kvist M, Natri A, Jarvinen M. Location and distribution of non-collagenous matrix proteins in musculoskeletal tissues of rat. Histochem J 1998;30:799–810.
36. Jozsa L, Lehto M, Kannus P, Kvist M, Reffy A, Vieno T, Jarvinen M, Demel S, Elek E. Fibronectin and laminin in Achilles tendon. Acta Orthop Scand 1989;60:469–471.
37. Mow VC, Mak AF, Lai WM, Rosenberg LC, Tang LH. Viscoelastic properties of proteoglycan subunits and aggregates in varying solution concentrations. J Biomech 1984;17:325–338.
38. Hooley CJ, Cohen RE. A model for the creep behaviour of tendon. Int J Biol Macromol 1979;1:123–132.
39. Kannus P. Structure of the tendon connective tissue. Scand J Med Sci Sports 2000;10:312–320.
40. Parry DA, Craig AS, Barnes GR. Tendon and ligament from the horse: an ultrastructural study of collagen fibrils and elastic fibers as a function of age. Proc R Soc London Ser B Biol Sci 1978;203:293–303.
41. Ross R. The elastic fiber. J Histochem Cytochem 1973;21: 199–208.
42. Minns RJ, Soden PD, Jackson DS. The role of the fibrous components and ground substance in the mechanical properties of biological tissues: a preliminary investigation. J Biomech 1973;6:153–165.
43. Gosline JM. The elastic properties of rubber-like proteins and highly extensible tissues. Symp Soc Exp Biol 1980;34:331–357.
44. Evans JH, Nachemson AL. Biomechanical study of human lumbar ligamentum flavum. J Anat 1969;105:188–189.
45. Agarwal BD, Broutman LJ. Analysis and performance of fiber composites New York: Wiley; 1980.
46. Gordon JE. The new science of stronmg materials. Harmondsworth: Penguin Books Ltd.; 1976.
47. Aspden RM. Fiber reinforcing by collagen in cartilage and soft connective tissues. Proc R Soc London Ser B Biol Sci 1994;B-258:195–200.
48. Goh KL, Aspden RM, Mathias KJ, Hukins DWL. Effect of fiber shape on the stresses within fibers in fiber-reinforced composite materials. Proc R Soc London Ser 1999;A-455: 3351–3361.
49. Goh KL, Mathias KJ, Aspden RM, Hukins DWL. Finite element analysis of the effect of fiber shape on stresses in an elastic fiber surrounded by a plastic matrix. J Mater Sci 2000;35:2493–2497.
50. Kelly A, Macmillan NH. Strong Solids. 3rd ed.; Oxford: Oxford University Press; 1986.
51. DeVente JE, Lester GE, Trotter JA, Dahners LE. Isolation of intact collagen fibrils from healing ligament [letter]. J Electron Microsc 1997;46:353–356.
52. Holmes DF, Chapman JA, Prockop DJ, Kadler KE. Growing tips of type-I collagen fibrils formed in vitro are near paraboloidal in shape, implying a reciprocal relationship between accretion and diameter. Proc Natl Acad Sci USA 1992;89: 9855–9859.
53. Kadler KE, Holmes DF, Trotter JA, Chapman JA. Collagen fibril formation. Biochem J 1996;316 (Pt 1): 1–11.
54. Aspden RM. The theory of fiber reinforced composite materials applied to changes in the mechanical properties of the cervix during pregnancy. J Theor Biol 1988;130:213–221.
55. Cabaud HE. Biomechanics of the anterior cruciate ligament. Clin Orthop 1983; 26–31.
56. Dodds JA, Arnoczky SP. Anatomy of the anterior cruciate ligament: a blueprint for repair and reconstruction. Arthroscopy 1994;10:132–139.
57. Noyes FR, DeLucas JL, Torvik PJ. Biomechanics of anterior cruciate ligament failure: an analysis of strain-rate sensitivity and mechanisms of failure in primates. J Bone Joint Surg Am 1974;56:236–253.
58. Hukins DWL, Kirby MC, Sikoryn TA, Aspden RM, Cox AJ. Comparison of structure, mechanical properties and functions of lumbar spinal ligaments. Spine 1990;15:787–795.
59. Viidik A. Functional properties of collagenous tissues. Int Rev Connect Tissue Res 1973;6:127–215.
60. Claes L, Neugebauer R. In vivo and in vitro investigation of the long-term behaviour and fatigue strength of carbon fiber ligament replacement. Clin Orthop 1985;186:99–111.
61. Tkaczuk H. Tensile properties of human lumbar longitudinal ligaments. Acta Orthop Scand Suppl 1968; 115.
62. Shah JS, Jayson MIV, Hampson WGJ. Mechanical implications of crimping in collagen fibers of human spinal ligaments. Proc Instn Mech Eng [H], J Eng Med 1979;8: 95–102.
63. Kirby MC, Sikoryn TA, Hukins DWL, Aspden RM. Structures and mechanical properties of the longitudinal ligaments and ligamenta flava of the spine. J Biomed Eng 1989;11:192–196.
64. Brand RA. Knee ligaments: a new view. J Biomech Eng 1986;108:106–110.
65. Viidik A. Tensile strength properties of Achilles tendon systems in trained and untrained rabbits. Acta Orthop Scand 1969;40:261–272.
66. Rigby J, Hirai N, Spikes JD, Eyring M. The mechanical properties of rat tail tendon. J Gen Physiol 2003;43:265–283.
67. Haut RC, Little RW. A constitutive equation for collagen fibers. J Biomech 1972;5:423–430.
68. Dorrington KL. The theory of viscoelasticity in biomaterials. Symp Soc Exp Biol 1980;34:289–314.
69. Cohen RE, Hooley CJ, McCrum NG. Viscoelastic creep of collagenous tissue. J Biomech 1976;9:175–184.
70. Kennedy JC, Hawkins RJ, Willis RB, Danylchuck KD. Tension studies of human knee ligaments. Yield point, ultimate failure, and disruption of the cruciate and tibial collateral ligaments. J Bone Joint Surg Am 1976;58:350–355.
71. Neumann P, Keller TS, Ekstrom L, Perry L, Hansson TH, Spengler DM. Mechanical properties of the human lumbar anterior longitudinal ligament. J Biomech 1992;25:1185–1194.
72. Sikoryn TA, Hukins DWL. Mechanism of failure of the ligamentum flavum of the spine during in vitro tensile tests. J Orthop Res 1990;8:586–591.
73. Azangwe G, Mathias KJ, Marshall D. Macro and microscopic examination of the ruptured surfaces of anterior cruciate ligaments of rabbits. J Bone Joint Surg Br 2000;82:450–456.
74. Barry D, Ahmed AM. Design and performance of a modified buckle transducer for the measurement of ligament tension. J Biomech Eng 1986;108:149–152.
75. Beynnon B, Howe JG, Pope MH, Johnson RJ, Fleming BC. The measurement of anterior cruciate ligament strain in vivo. Int Orthop 1992;16:1–12.
76. Woo SL, Newton PO, MacKenna DA, Lyon RM. A comparative evaluation of the mechanical properties of the rabbit

medial collateral and anterior cruciate ligaments. J Biomech 1992;25:377–386.

77. Lam TC, Shrive NG, Frank CB. Variations in rupture site and surface strains at failure in the maturing rabbit medial collateral ligament. J Biomech Eng 1995;117:455–461.
78. Sikoryn TA, Hukins DWL. Failure of the longitudinal ligaments of the spine. J Mater Sci Lett 1988;7:1345–1349.
79. Gupta P, Subramanian KN, Brinker WO, Gupta AN. Tensile strength of canine cruciate ligaments inthe dog. Am J Vet Res 1971;32:183–190.
80. Njus GO, Njus NM. A non-contact method for determining cross-sectional area of soft tissue. Trans Orthopaed Res Soc 1984;32:126–131.
81. Woo SL, Danto MI, Ohland KJ, Lee TQ, Newton PO. The use of a laser micrometer system to determine the cross-sectional shape and area of ligaments: a comparative study with two existing methods. J Biomech Eng 1990;112:426–431.
82. King GJ, Pillon CL, Johnson JA. Effect of in vitro testing over extended periods on the low-load mechanical behaviour of dense connective tissues. J Orthop Res 2000;18:678–681.
83. Kwan MK, Lin TH, Woo SL. On the viscoelastic properties of the anteromedial bundle of the anterior cruciate ligament. J Biomech 1993;26:447–452.
84. Provenzano P, Lakes R, Keenan T, Vanderby, Jr. R. Nonlinear ligament viscoelasticity. Ann Biomed Eng 2001;29: 908–914.
85. Silver FH, Christiansen DL, Snowhill PB, Chen Y. Role of storage on changes in the mechanical properties of tendon and self-assembled collagen fibers. Connect Tissue Res 2000; 41:155–164.
86. Woo SL, Gomez MA, Akeson WH. The time and history-dependent viscoelastic properties of the canine medical collateral ligament. J Biomech Eng 1981;103:293–298.
87. Thornton GM, Oliynyk A, Frank CB, Shrive NG. Ligament creep cannot be predicted from stress relaxation at low stress: a biomechanical study of the rabbit medial collateral ligament. J Orthop Res 1997;15:652–656.
88. Thornton GM, Shrive NG, Frank CB. Altering ligament water content affects ligament pre-stress and creep behaviour. J Orthop Res 2001;19:845–851.
89. Speer KP, Warren RF, Wickiewicz TL, Horowitz L, Henderson L. Observations on the injury mechanism of anterior cruciate ligament tears in skiers. Am J Sports Med 1995;23:77–81.
90. Fischer JF, Leyvraz PF, Bally A. A dynamic analysis of knee ligament injuries in alpine skiing. Acta Orthop Belg 1994;60:194–203.
91. Emerson RJ. Basketball knee injuries and the anterior cruciate ligament. Clin Sports Med 1993;12:317–328.
92. Azangwe G, Mathias KJ, Marshall D. The effect of torsion on the appearance of the rupture surface of the ACL of rabbits. Knee 2002;9:31–39.
93. Woo SL, Orlando CA, Gomez MA, Frank CB, Akeson WH. Tensile properties of the medial collateral ligament as a function of age. J Orthop Res 1986;4:133–141.
94. Davison PF. The organization of collagen in growing tensile tissues. Connect Tissue Res 1992;28:171.
95. Greenlee, Jr. TK, Pike D. Studies on tendon healing in the rat. J Plast Reconstr Surg 1071;48:260–270.
96. Matyas JR, Anton MG, Shrive NG, Frank CB. Stress governs tissue phenotype at the femoral insertion of the rabbit MCL. J Biomech 1995;28:147–157.
97. Moriggl B, Kumai T, Milz S, Benjamin M. The structure and histopathology of the “enthesis organ” at the navicular insertion of the tendon of tibialis posterior. J Rheumatol 2003;30:508–517.
98. Cotton FJ, Morrison GM. Artifical ligaments at the knee: a technique. N Engl J Med 1934;210:1331–1332.
99. Fujikawa K, Kobayashi T, Sasazaki Y, Matsumoto H, Seedhom BB. Anterior cruciate ligament reconstruction with the Leeds-Keio artificial ligament. J Long Term Eff Med Implants 2000;10:225–238.
100. Chen CH, Chen WJ, Shih CH. Arthroscopic reconstruction of the posterior cruciate ligament: a comparison of quadriceps tendon autograft and quadruple hamstring tendon graft. Arthroscopy 2002;18:603–612.
101. Ruiz AL, Kelly M, Nutton RW. Arthroscopic ACL reconstruction: a 5-9 year follow-up. Knee 2002;9:197–200.
102. Schroven IT, Geens S, Beckers L, Lagrange W, Fabry G. Experience with the Leeds-Keio artificial ligament for anterior cruciate ligament reconstruction. Knee Surg Sports Traumatol Arthrosc 1994;2:214–218.
103. Drogset JO, Grontvedt T. Anterior cruciate ligament reconstruction with and without a ligament augmentation device: results at 8-Year follow-up. Am J Sports Med 2002;30:851–856.
104. Friedman MJ, Sherman OH, Fox JM, Del PW, Snyder SJ. Ferkel RJ. Autogeneic anterior cruciate ligament (ACL) anterior reconstruction of the knee. A review. Clin Orthop Relat Res 1985;9–14.
105. Alexander H, Weiss AB. Editorial Comment. Clin Orthop Relat Res 1985;2–3.
106. Fujikawa K, Ohtani T, Matsumoto H, Seedhom BB. Reconstruction of the extensor apparatus of the knee with the Leeds-Keio ligament. J Bone Joint Surg Br 1994;76:200–203.
107. Jenkins DH, Forster IW, McKibbin B, Ralis ZA. Induction of tendon and ligament formation by carbon implants. J Bone Joint Surg Br 1977;59:53–57.
108. Turner IG, Thomas NP. Comparative analysis of four types of synthetic anterior cruciate ligament replacement in the goat: in vivo histological and mechanical findings. Biomaterials 1990;11:321–329.
109. Marumo K, Kumagae Y, Tanaka T, Fujii K. Long-term results of anterior cruciate ligament reconstruction using semitendinosus and gracilis tendons with Kennedy ligament augmentation device compared with patellar tendon autografts. J Long Term Eff Med Implants 2000;10:251–265.
110. Fukubayashi T, Ikeda K. Follow-up study of Gore-Tex artificial ligament—special emphasis on tunnel osteolysis. J Long Term Eff Med Implants 2000;10:267–277.
111. Woo SL, Hildebrand K, Watanabe N, Fenwick JA, Papageorgiou CD, Wang JH. Tissue engineering of ligament and tendon healing. Clin Orthop 1999; S312–S323.
112. Huard J, Li Y, Peng H, Fu FH. Gene therapy and tissue engineering for sports medicine. J Gene Med 2003;5:93–108.
113. Gentleman E, Lay AN, Dickerson DA, Nauman EA, Livesay GA, Dee KC. Mechanical characterization of collagen fibers and scaffolds for tissue engineering. Biomaterials 2003;24: 3805–3813.
114. DesRosiers EA, Yahia L, Rivard CH. Proliferative and matrix synthesis response of canine anterior cruciate ligament fibroblasts submitted to combined growth factors. J Orthop Res 1996;14:200–208.
115. Martinek V, Latterman C, Usas A, Abramowitch S, Woo SL, Fu FH, Huard J. Enhancement of tendon-bone integration of anterior cruciate ligament grafts with bone morphogenetic protein-2 gene transfer: a histological and biomechanical study. J Bone Joint Surg Am 2002;84-A:1123–1131.

Further Reading

Akeson WH, Pedowitz R, O‘Connor JJ, editors. Knee Ligaments: Structure, Function, Injury and Repair. 2nd ed. New York: Lippincott Williams & Wilkins; 2003.

Mow VC, Hayes WC, editors. Basic Orthopaedic Biomechanics. New York: Raven Press; 1991.

Evans CH, Scully SP, guest editors. Orthopaedic Gene Therapy. Clinical Orthopaedics and Related Research. Volume 379 Suppl., 2000.

See also Bone and teeth, properties of; cartilage and meniscus, properties of.

LINEAR VARIABLE DIFFERENTIAL TRANSFORMERS

Sunil Kesavan
Akebono Corporation
Farmington Hills, Michigan

Narender Reddy
The University of Akron
Akron, Ohio

INTRODUCTION

Several methods of transduction are available to convert physiological events into electrical signals. Basic physiological variables are first converted by sensing elements into variables that can easily be measured by available transducers. One such transducer, the linear variable differential transformer, commonly abbreviated as LVDT (some manufacturers designate it as LDVT — linear differential voltage transformer), is used to convert mechanical displacement into proportional electronic signals. LVDTs are capable of measuring physiological variables, such as displacement, pressure, force, and acceleration, which are either available in the form of a linear displacement or can be converted into such movement.

THEORY

An LVDT is an inductive electromechanical transducer that uses a primary (energizing) coil and two series-opposed secondary coils. This mode of connecting the secondaries serves to mutually cancel out the secondary voltages. In this popular configuration, due to Shaevitz (1), the primary winding is symmetrically placed with respect to the secondary windings on a cylindrical former. The former surrounds a free-moving rod-shaped magnetic core, which provides a path for the magnetic flux linking the coils (Fig. 1). The magnetic core is connected to a sensing device like a movable diaphragm. Movement of the sensor induces core movement, which in turn produces voltage variations that are measured directly.

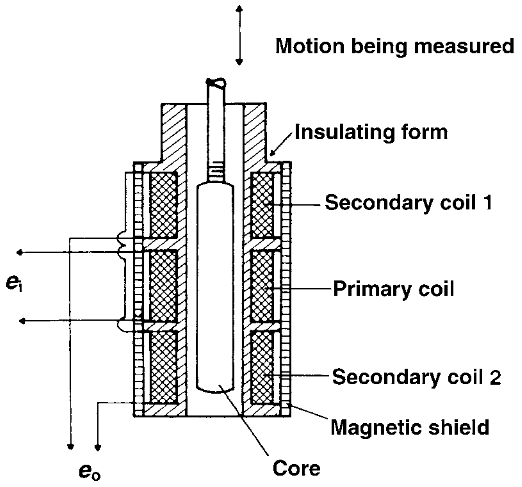


Figure 1. Schematic of the cutaway view of an LVDT showing the core, primary coil, and two secondary coils: (e_i) excitation voltage; (e_0) output voltage.

When the sliding magnetic core is in the central (null) position, the electromotive forces (emfs) generated in the secondaries are equal, and the net output voltage, e_0 is, therefore, zero. Movement of the core from this central position causes the mutual inductance (coupling) for one coil to increase and the other coil to decrease. The amplitude of the output voltage, e_0, being the difference between the emfs in the two secondaries, varies approximately linearly with the position of the core on either side of the null position (Fig. 2). The differential secondary connection in the LVDT causes the phase of the output voltage to change by 180° as the core passes through the null position. The output voltage, e_0, is generally out of phase with the excitation voltage, e_i. The phase shift is dependent on the frequency of e_i, and each LVDT has a particular frequency at which phase shift is zero.

FABRICATION

The LVDT features essentially frictionless measurement and long mechanical life, because there is no mechanical

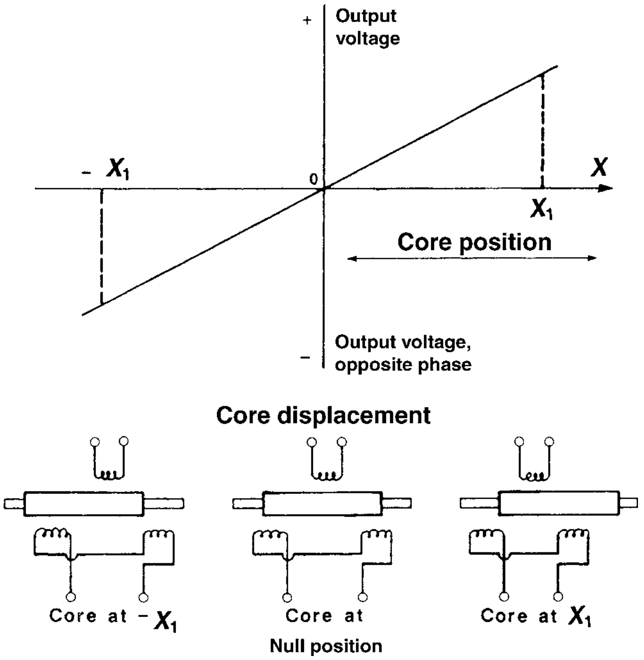


Figure 2. Output voltage of an LVDT as a linear function of core position.

contact between the enclosed coil assembly and the separate freely moving core within the coil assembly (Fig. 1).

A typical alternating current (ac) LVDT core consists of a uniformly dense cylindrical slug of a high permeability nickel–iron alloy. The core is internally threaded to accept nonmagnetic core rods from an external sensing or actuating element. The core moves within a cylindrical coil assembly. Hollow cores are employed when a low mass core is desired. Recently, researchers have developed lightweight glass-covered amorphous wire cores that can be used to fabricate high sensitivity LVDTs with good mechanical and corrosion resistance (2). The primary and secondary windings are spaced symmetrically by winding them on a slotted cylindrical former. To avoid material corrosion or insulation leakage, the windings are impregnated with an insulating varnish under a vacuum. The coils are encapsulated in epoxy for further mechanical and moisture protection. Magnetic or ferromagnetic materials in the proximity of an LVDT can disrupt its magnetic field. The magnetic field of an LVDT may also induce eddy currents into nonmagnetic materials in its vicinity. These currents in turn would create a magnetic flux that would interfere with the LVDT output. These problems are avoided in practice by enclosing the LVDT in a case fabricated from an alloy, a high permeability iron, or a stainless steel. The LVDT assembly is then mounted in a C- or split block. LVDTs that can measure rotational movement are also available.

In addition to the ac LVDTs described in the previous sections, direct current (dc) LVDTs are also available (3,4). These LVDTs, in addition to having all the advantages of ac LVDTs, possess the simplicity of dc operation. They consist of two integral parts: an ac-operated LVDT and a carrier generator–signal conditioning module. The small carrier system eliminates the need for the ac excitation, demodulation, and amplification equipment required for conventional ac LVDTs. This cuts down the cost and reduces the volume of LVDT instrumentation; dc units can be battery operated or be supplied by a simple dc power supply (3,4). Also, any dc meter can be employed to read the LVDT output. These advantages, coupled with the small size of the dc LVDTs, make them attractive for use in hospitals and other medical environments.

The LVDTs have several advantages and a few disadvantages (3–6) as briefly reviewed next.

1. Essentially frictionless operation and long mechanical life: As described in the previous section, the LVDT has no moving mechanical contact between the moving core and the windings. This ensures that LVDTs have a fast dynamic response as no additional load apart from the core mass is imposed on the measured event. In addition, this helps LVDTs to have a long, essentially infinite, mechanical life.
2. Good in hostile environments: LVDTs can be manufactured to withstand the vagaries of chemical corrosion and extremes of temperature and pressure. This is facilitated by the separation between the core and windings of the LVDT. Only a static seal is required to isolate the coil assembly from hostile environments.
3. Extremely high resolution: LVDTs can respond to extremely small displacements. Microdisplacement LVDT transducers capable of measuring displacements down to 100 pm have been fabricated (7).
4. Null repeatability: The null position of an LVDT is very repeatable, even with large temperature variations.
5. Input–Output isolation: Since the primary and secondary windings are isolated from each other, the signal ground can be isolated from the excitation ground.
6. Cross-axis rejection: The LVDT is only responsive to axial core motion. Cross-axis motion induced by conditions such as jarring or continuous vibration will not affect the LVDT output.
7. Overtravel damage resistance: As the LVDT core can pass completely through the coil assembly, the transducer is inherently immune to damage from unanticipated overtravel that can be encountered in applications where materials or structures can yield or fail.
8. Absolute output: Unlike a lot of other transducers that are incremental output devices, LVDTs are absolute output devices, that is, the displacement information from an LVDT is not lost if the system loses power. When the measuring system is restarted, the LVDTs output value will be the same as it was before the power failure occurred.

All these advantages, in addition to their reasonable cost, have made the LVDT an attractive displacement measurement technique. However, LVDTs for use in medical applications have the following disadvantages: (1) They require a constant amplitude excitation of high frequency. (2) They cannot be used in the vicinity of equipment that creates strong magnetic fields.

LVDT INSTRUMENTATION

Instrumentation normally used with an ac LVDT should perform the following functions (3,4).

Excitation

An LVDT needs an ac input of constant amplitude at a frequency that is not readily available. Hence, an oscillator of the appropriate frequency has to be connected to an amplifier with amplitude regulation on its output.

Amplification

As in the case of most transducers, the low level outputs of LVDTs require amplification. One procedure for amplification employs two steps: (1) use of an ac carrier–amplifier before demodulation; and (2) a dc amplifier after the demodulator (3,4).

Demodulation

As discussed earlier, the output of an LVDT remains proportional to the displacement while it undergoes a

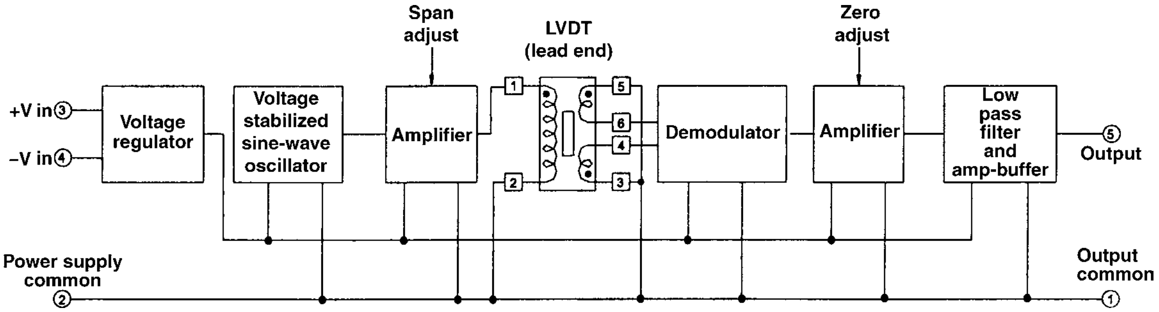


Figure 3. Block diagram for an LVDT employing a series 1000 oscillator–demodulator supplied by Trans-Tek, Inc. (Courtesy of Trans-Tek Inc.)

phase shift of 180° when the core goes through the null. When this LVDT output is connected to a voltmeter, the meter will register the same reading for equal amounts of core displacement on either side of the null position. This lack of directional sensitivity has to be overcome if one has to tell to which side of the null the core is displaced. Two techniques can be used to confer directional sensitivity on an LVDT output. In one technique, the core is offset, and the operation is centered on a position other than the null point. In this case, the output signal either increases or decreases. The other procedure uses a phase-sensitive demodulator (also called a detector). Several such devices are available and are discussed in detail elsewhere (8). The simplest forms employ diode rectification while the complex forms involve synchronous demodulation. Figure 3 shows the block diagram for an LVDT employing a series 1000 oscillator–demodulator supplied by Trans-Tek, Inc. (9).

The demodulator confers directional sensitivity on its input (output of the LVDT), which is either in phase with, or 180° out of phase with, the carrier signal (10). The demodulator output e_0 is usually sent to a low pass filter that will pass only the frequencies present in x and reject all higher frequencies created by the modulation procedure. Obviously, demodulation is not required if the LVDT transducer is to be used only on one side of the null position.

Recent developments allow all LVDT support circuitry to be accomplished using an inexpensive flexible field programmable analog array (FPAA). The FPAA consists of “configurable analog blocks” consisting of switched-capacitor op-amp cells surrounded by a programmable interconnect and I/O structure (11).

dc Power

Stable dc voltage sources are required for operation of the electronics associated with LVDTs. The dc LVDTs available at the present time employ a microcircuit module including all the electronics needed to provide ac excitation to the primary of the LVDT and to demodulate and amplify the analog LVDT signal. The module is mounted in tandem with the LVDT and only increases the effective LVDT length slightly.

Figure 4 shows the block diagram for a dc LVDT (9). The oscillator produces a constant amplitude sine wave excitation for the primary of the LVDT. A phase sensitive demodulator and an RC filter network process the secondary coil output. Some dc LVDT modules are furnished with a reverse polarity protector for the dc power input. dc LVDTs are becoming increasingly popular due to their advantages in the areas of calibration and signal conditioning.

SELECTION CRITERIA

Several criteria have to be considered in selecting a particular LVDT for a certain application (12). The manufacturer supplies several of these parameters as specification criteria.

Total Stroke

Stroke-length specification in the selection of an LVDT for a particular application is governed by the displacement to be measured. LVDTs can be custom-made for either short- (up to 0.01 m) or long-stroke (up to 1.5 m) operation; however, cost of fabrication increases greatly with increase in length, and lengths over 0.03 m may not be cost effective.

Linearity and the Nominal Linear Range

The output of an LVDT is a nearly linear function of core position for a rather wide range on either side of the balance (null) position (Fig. 2). A nominal linear range is defined for an LVDT as the core displacement on either side of the balance position for which the LVDT output as a function of displacement remains a straight line. Outside this range, the output starts to deviate gradually from the ideal straight line in the form of a smooth curve. Linearity of an LVDT is defined as “the maximum deviation from a best-fit straight line (applied to a plot of LVDT output voltage vs. core displacement) within the nominal linear

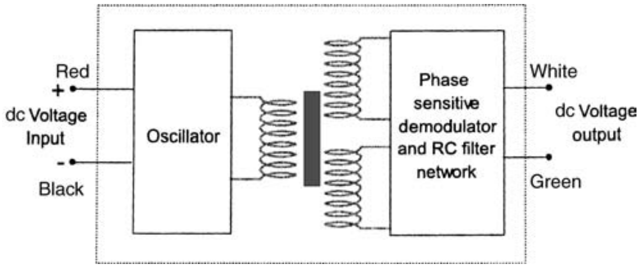


Figure 4. Block diagram of a dc LVDT. (Courtesy of Trans-Tek Inc.)

range" (12). Linearity is usually expressed as a percentage of the full scale. A typical LVDT has a linearity of about ±0.25%.

Sensitivity

The sensitivity of an LVDT is usually expressed as the output in millivolts (or V) per 0.001 m core displacement per volt input. Normally, both the input voltage and the frequency are specified as well, because voltage sensitivity may vary with frequency over a limited frequency range. A typical miniature LVDT transducer has a sensitivity of ~8–200 mV out/0.001 m/V input.

Resolution

Resolution of an LVDT is the smallest core movement that can produce an observable change in the voltage output (12). With careful circuit design, displacements smaller than 100 nm can be detected.

Armature Mass

The mass of the armature (core) of the LVDT should be small so as not to unduly load the measured event. A reduction in the length of the LVDT results in a reduction in either the linearity or the maximum linear range, whereas sensitivity increases.

Excitation Frequency and Voltage

The sensitivity of the LVDT depends on both the excitation voltage and frequency. Normally, a sinusoidal voltage of 3–15 V rms amplitude and a frequency of 60 Hz–20 kHz is used for the excitation of LVDTs. The sensitivity of an LVDT increases with the excitation frequency, particularly at the lower part of the operating frequency range (12). Normally, an excitation frequency range of 1–5 kHz produces optimal LVDT operation.

Operating Environment

LVDTs have the advantage of being available in hostile-environment-proof format. Transducers designed to withstand both high and cryogenic temperatures and high pressures are available. Immersion-type LVDTs resistant to corrosive liquids are also available. Normally, specification criteria for an LVDT include information on the temperature range of operation and the temperature coefficient.

Residual Voltage Output

The residual voltage output is the LVDT output when the core is in the null position. This should ideally be zero; however, the null voltages and the harmonics of the excitation source do not cancel, resulting in a nonzero residual output (12). In practice, the residual voltage is about 1% of that obtained with maximum displacement.

Repeatability

Repeatability, the ability of the LVDTs to give the same output if the core is displaced and returned to the original position is an important consideration. LVDTs with repeatability better than 100 nm are available for some critical applications.

MEDICAL APPLICATIONS

LVDTs are used in medical applications and research to measure physiological variables that are either available in the form of a linear displacement or can be converted into such movement. LVDTs for medical applications can be readily fabricated in very small sizes with low mass cores. This will ensure that only a negligible force is imposed on the measured physiological event. Also, due to the low alternating currents in the windings, negligible magnetic load is imposed. When not in use, the core remains in the null position, and no force is imposed on the measured event. Even when the core is displaced from null, the load imposed on the event is small. These advantages, coupled with the general advantages of LVDTs discussed in the previous paragraphs, make these transducers very attractive for physiological measurements.

One early application of LVDTs was in the fabrication of invasive blood pressure measurement transducers (13). These transducers consisted of three essential parts: (1) a dome with pressure fittings, (2) a stainless steel diaphragm and core assembly, and (3) the LVDT coils. Pressure transmitted via the catheter exerts a force on the diaphragm. This causes a movement of the diaphragm, which in turn manifests itself in a movement of the core attached to it. Movement of the core of the LVDT creates a proportional output that can then be recorded after suitable electronic circuit processing. Catheter tip and implantable transducers employed the same principle (13). However, these rugged LVDT blood pressure transducers have been supplanted by cost-effective microelectromechanical system (MEMS) type transducers (14).

Another application for LVDT transducers is in indentation tests on tissue to determine mechanical properties. The authors have developed an LVDT indenter for the characterization of the mechanical properties of skin and the underlying soft tissue (Fig. 5). The indenter uses a loaded hemispherical tip coupled with a load cell-LVDT system for simultaneously measuring both the force and displacement during indentation tests. This information in turn was used to evaluate soft tissue properties. Walsh and Schettini (15) used a similar indenter to measure the *in vivo* viscoelastic response properties of brain tissue. Oculotonometers operating on the same principle and designed to indent the corneoscleral shell use LVDTs to measure deflections in the micrometer range (16). In another similar application, Gunner et al. (17) used an LVDT transducer-mounted extensometer to measure the *in vivo* recoil characteristics of human skin. The device consisted of two flat rectangular tabs, one fixed and the other capable of rectilinear sliding motion, attached to the test skin surface with double-sided adhesive tape. This combination was attached to an LVDT displacement transducer. Behavior of the skin resulting from the movement of the tabs was converted by the LVDT into electronic signals that were then analyzed to characterize the skin.

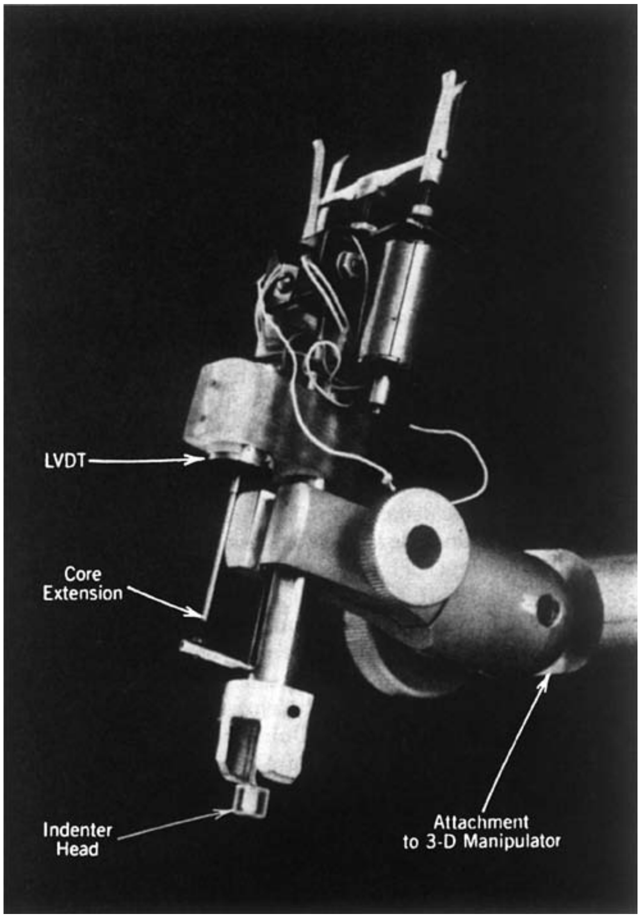


Figure 5. An LVDT skin and tissue indenter. (Courtesy of N. P. Reddy, J. Kagan and G. V. B. Cochran.)

Christiansen et al. (18) used a similar device for the viscoelastic characterization of skin. These examples are just a few of the myriad potential applications for LVDT transducers in soft tissue characterization.

Several radiological and neurological devices have incorporated LVDTs. For example, Laser Diagnostic Technologies of San Diego, CA, incorporated a position-sensing DC-DC LVDT into a scanning laser tomography instrument designed for retinal topography (9). The stable and repeatable DCDT output is part of a continuous feedback loop in the scanner's on-board logic control system. Radionics, Inc., employed an LVDT in a sophisticated modular probe drive used to support the precise implantation of deep brain stimulating electrodes (9). The device uses a push–pull cable drive mechanism to move the carrier that guides the probe to the desired location in the brain. The LVDT is mounted at the top of the mechanism and is used to accurately monitor probe position (Fig. 6). The LVDT used in this application was sealed to resist moisture and was modified to withstand the rigors of steam sterilization.

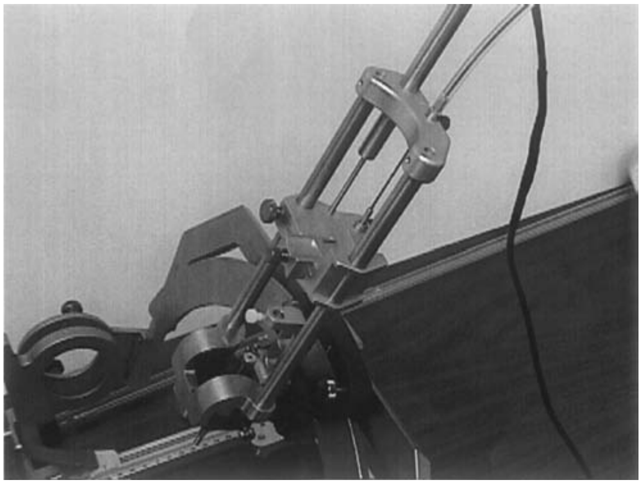


Figure 6. Mechanism of modular brain probe drive showing the LVDT used to help in precise electrode placement. [Courtesy of Radionics (a division of Tyco Healthcare).]

LVDTs have been used in endocrinology and pharmacology to evaluate *in vitro* and *in vivo* contractile properties of vascular smooth muscle. Erdos et al. (19) designed an *in vitro* isotonic myograph employing an LVDT. The device resembled a beam-type balance. One arm of the device was connected to the contracting muscle specimen, and the other arm was counterbalanced by a suspended weight. Motion of the weight was translated via the movement of an LVDT core into electronic signals.

Gow (20) employed a novel LVDT electronic caliper for the continuous monitoring of arterial diameter changes during pulsation. This low mass device was found to possess a rapid response time with a natural resonant frequency greater than 180 Hz. Shykoff et al. (21) used LVDT measurements of changes in diameter of dorsal hand veins to establish diameter, pressure, and compliance relationships. LVDT-based devices have been used to evaluate the *in vivo* vascular effects of drugs with the dorsal hand vein technique (22). For example, Landau et al. (23) used an LVDT-based device to evaluate the magnesium-induced vasodilation in the dorsal hand vein. A similar technique was used by Streeten and Anderson (24) to measure venous contractile responses to locally infused norepinephrine.

LVDTs have been used in numerous orthopedic and dental devices. Chen et al. (25) used an LVDT bonded to the mandibular first molars to quantify mandibular deformation during mouth opening. Other possible medical applications are the mapping of facial contours before and after maxillofacial surgery and the profiling of spinal deformation in abnormalities like scoliosis. Buhler et al. (26) and Flamme et al. (27) used LVDTs to quantify micromotion in orthopedic implants. Recently, Dong et al.(28) incorporated an LVDT into a device for quantitative assessment of tension in wires of fine-wire external orthopedic fixators.

An LVDT was used to calibrate finger movements and to correlate these movements with surface electromyograms from the flexor digitorum superficialis muscles in the forearm (29,30). This work was designed to develop techniques for control of anthropomorphic teleoperator fingers using surface electromyographic signals obtained from the forearm.

Wang et al. (31) used six spring-loaded LVDTs in an experimental technique to measure three-dimensional (3D), six-degrees-of-freedom motion of human joints. Rotary LVDTs are useful for measuring joint angles. For measuring 3D rotations, rotary LVDTs are incorporated into six-degree-of-freedom motion linkages.

As illustrated in the applications discussed above, LVDTs are highly suited for biomedical device and research applications requiring accurate displacement measurements with high resolution, input–output isolation, and cross-axis rejection. Although LVDTs are being replaced by miniaturized, cost-effective transducers utilizing advanced fabrication technologies in many applications, their advantages still render them excellent candidates for biomedical applications.

ACKNOWLEDGMENT

The authors would like to acknowledge the help provided by Rema Menon in the preparation of this manuscript.

BIBLIOGRAPHY

1. Schaevitz H. The linear variable differential transformer. Proc Soc Stress Anal 1947;4:79–88.
2. Chiriac H, Hristoforou E, Neagu M, Pieptanariu M. Linear variable differential transformer sensor using glass-covered amorphous wires as active core. J Magn Magn Mater 2000; 215:759–761.
3. Schaevitz Engineering. Technical bulletins 1002D and 7007. Pennsauken (NJ) 1986.
4. Schaevitz Sensors. Shaevitz Sensor Solutions. Catalog No. SCH-2001 Hampton (VA) 2000.
5. [Anonymous]. No date. An LVDT Primer. [Online]. Macro Sensors. Available at www.macrosensors.com. 2005 March 8.
6. Weinstein E. LVDTs on the factory floor. Instrum Control Syst 1982;55:59–61.
7. Sydenham PH. Microdisplacement transducers. J Phys E 1972;6:721–733.
8. Szczyrbak J, Schmidt EDD. LVDT signal conditioning techniques. Meas Control 1997;183:103–111.
9. [Anonymous]. No date. LVDT application. [Online]. Trans-Tek Inc. Available at www.transtekinc.com. 2005 March 8.
10. Doebelin EO. Measurement Systems: Application and Design. 4th ed. New York: McGraw-Hill; 1990.
11. Severn J, October. 2001. New analog interface for LVDTs. [Online]. Industrial Technology. Available at www.industrialtechnology.co.uk. 2005 March 8.
12. Anonymous, Finding the right LVDT. Instrum Control Syst 1977;50:61–62.
13. [Anonymous]. No date. LVDT Applications. [Online]. Macro Sensors. Available at www.macrosensors.com. 2005 March 8.
14. Seeley RS. 1996. The future of medical microelectromechanical systems. [Online]. Available at www.devicelink.com/mem/archive/96/01/003.html. 2005 March 8.
15. Walsh EK, Schettini A. A pressure-displacement transducer for measuring brain tissue properties *in vivo*. J Appl Physiol 1975;38:187–189.
16. Stepanik J. The Mackay-Marg Tonometer. Acta Ophthal 1970;48:1140.
17. Gunner CW, Hutton WC, Burlin TE. An apparatus for measuring the recoil characteristics of human skin *in vivo*. Med Biol Eng Comput 1979;17:142–144.
18. Christiansen MS, Hargens III CW, Nacht S, Gans EH. Viscoelastic properties of intact human skin: Instrumentation, hydration effects, and the contribution of the stratum corneum. J Invest Dermatol 1977;69:282–286.
19. Erdos EG, Jackman V, Barnes WC. Instrument for recording isotonic contractions of smooth muscles. J Appl Physiol 1962;17: 307–308.
20. Gow BS. An electrical caliper for measurement of pulsatile arterial diameter changes *in vivo*. J Appl Physiol 1966;21: 1122–1126.
21. Shykoff BE, Hawari FI, Izzo JL. Diameter, pressure and compliance relationships in dorsal hand veins. Vasc Med 2001;6(2):97–102.
22. Pang YC. Autonomic control of venous system in health and disease: effect of drugs. Pharmacol Therapeut 2001;90:179–230.
23. Landau R, Scott JA, Smiley RM. Magnesium-induced vasodilation in the dorsal vein. BJOG (Bri J Obst Gyn) 2004;111: 446–451.
24. Streeten DHP, Anderson GH. Mechanisms of orthostatic hypotension and tachycardia in patients with pheochromocytoma. AJH 1996;9:760–769.
25. Chen DC, Lai YL, Chi LY, Lee SY. Contributing factors of mandibular deformation during mouth opening. J Dent 2000;28(8):583–588.
26. Buhler DW, Oxland TR, Nolte LP. Design and evaluation of a device for measuring three-dimensional motions of press-fit femoral stem prosthesis. Med Eng Phys 1999;19:187–199.
27. Flamme CH, Kohn D, Kirsch L, Hurschler C. Primary stability of different implants used in conjunction with high tibial osteomy. Arch Orth Trauma Surg 1999;119:450–455.
28. Dong Y, Saleh M, Yang L. Quantitative assessment of tension in wires of fine-wire external fixators. Med Eng Phys 2005;27:63–66.
29. Gupta V, Reddy NP. Surface electromyogram for the control of anthropometric teleoperator fingers. Weghorst SJ, Soeburg HB, Morgan KS, editors. Medicine Meets Virtual Reality: Healthcare in the Information Age. Amsterdam: IOP Press; 1996.
30. Devavaram A, Reddy NP. Intelligent systems for control of telemanipulators using surface EMG signals, submitted for publication.
31. Wang M, Bryant JT, Dumas GA. A new *in vitro* measurement technique for small three-dimensional joint motion and its application to the sacroiliac joint. Med Eng Phys 1996;18(6): 495–501.

Further Reading

Herceg ED. Handbook of Measurement and Control, Pennsauken (NJ): Schaevitz Engineering; 1972.

Anonymous. LVDTs remain "State-of-the-art. Meas Inspect Technol 1982;4(2):13–16.

Anonymous. Displacement transducers (linear variable differential transformer products review). Control Instrum 1984;16(8): 23–25.

Geddes LA, Baker LE. Principles of Applied Biomedical Instrumentation. 3rd ed. New York: Wiley-Interscience; 1989.

Webster JG. Medical Instrumentation: Application and Design. 3rd ed. New York: John Wiley & Sons; 1998.

See also BLOOD PRESSURE MEASUREMENT; INTEGRATED CIRCUIT TEMPERATURE SENSOR.

LITERATURE, MEDICAL PHYSICS. See MEDICAL PHYSICS LITERATURE.

LITHOTRIPSY

ALON Z. WEIZER
GLENN M. PREMINGER
Duke University Medical Center
Durham, North Carolina

INTRODUCTION

The clinical introduction of shock wave lithotripsy (SWL) by Chaussy in 1980 has revolutionized the way in which patients with renal and ureteral calculi are treated. Shock wave lithotripsy is a noninvasive method of fragmenting stones located inside the urinary tract. Since its initial introduction, SWL technology has advanced rapidly in terms of the means for shock wave generation, shock wave focusing, patient coupling, and stone localization. Despite rapid technological advances, most current commercial lithotripters are fundamentally the same; they produce a similar pressure waveform at the focus, which can be characterized by a leading shock front with a compressive wave followed by a trailing tensile wave (Fig. 1). The acoustic fields produced by different lithotripters differ from each other in terms of the peak amplitudes of the pressure waveform, pulse duration, beam size, total acoustic energy, and therefore, their overall performance.

Clinical experience has guided the technical development of second and third generation lithotripters with the aim of providing user convenience and multifunctionality of the device, rather than on further understanding of how SWL fragments calculi or injures surrounding renal tissue. Furthermore, the evolution of lithotripter design thus far has overwhelmingly relied upon the importance of the compressive wave component of the shock wave (positive portion of the sound wave), with almost total neglect of the contribution of the tensile component of the waveform. Consequently, current lithotripters have suffered from inferior fragmentation rates compared to the original HM3 lithotripter.

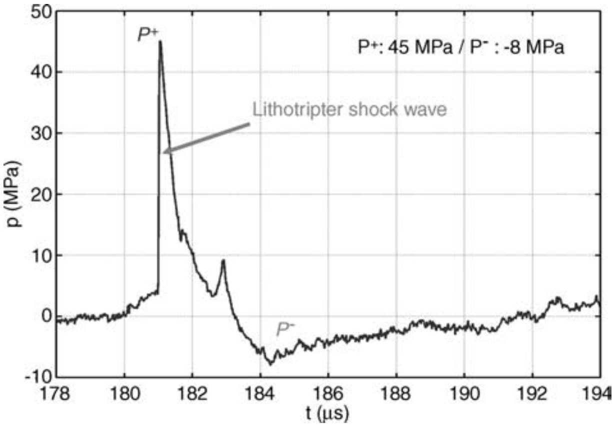


Figure 1. Pressure–time relationship of typical shock waves. Typical pattern of a lithotripter-generated shock wave. The shock wave is characterized by a leading positive or compressive component (P+) followed by a negative or tensile component (P−).

In contrast, significant progress in SWL basic science research has been made in the past 5 years to improve our understanding of the primary mechanisms for stone comminution (fragmentation) and tissue injury. It is now recognized that the disintegration of renal calculi in a lithotripter field is the consequence of dynamic and synergistic interaction of two fundamental mechanisms: stress wave-induced dynamic fracture in the form of nucleation, growth, and coalescence of preexisting microcracks inside the stone (1) and cavitation erosion caused by the violent collapse of bubbles near the stone surface (2,3). Similarly, two different mechanisms have been proposed for SWL-induced tissue injury: shear stress due to shock front distortion (4) and cavitation induced inside blood vessels, especially the expansion of intraluminal bubbles (5).

To understand how SWL fragments stones and causes tissue injury, the basic components of current lithotripters, the mechanisms behind stone fragmentation and kidney injury, and clinical results of the original electrohydraulic lithotripter in fragmenting kidney stones in patients will be described. In addition, the future directions of SWL will be reviewed based on current research that is investigating ways to make lithotripters more efficient and safer.

HISTORY AND EVOLUTION OF SWL

Physicists at Dornier Systems, Ltd. and Friedrich Shafen, Germany began experimenting with shock waves and their travel through water and tissue in 1963. Throughout the 1970s, numerous experimental lithotripters were developed that used new methods of shock wave generation and focusing as well as different techniques of stone localization. In addition, experimental studies were being performed *in vitro* and *in vivo* (in animal models) examining the effects of shock waves on various organs and tissues.

In 1980, Chaussy and associates successfully treated the first human and reported their first series of 72 patients in 1982 (6). Subsequently, $>$ 1800 articles have been published in the peer-reviewed literature, detailing the use of SWL for the management of renal and ureteral calculi. Moreover, numerous second and third generation devices have been introduced and are currently being used throughout the world. To understand how SWL results in stone fragmentation, the fundamentals of this technology are reviewed.

SWL PRINCIPLES

Despite the tremendous number of lithotripters currently available for fragmentation of renal and ureteral stones, all of these devices rely on the same laws of acoustic physics. Shock waves (i.e., high pressure sound waves) consisting of a sharp peak in positive pressure followed by a trailing negative wave are generated extracorporeally and passed through the body to fragment stones. Sounds waves readily propagate from a water bath or water medium into the human body, due to similar acoustic impedances.

As a consequence, all lithotripters share four main features: an energy source to generate the shock wave, a

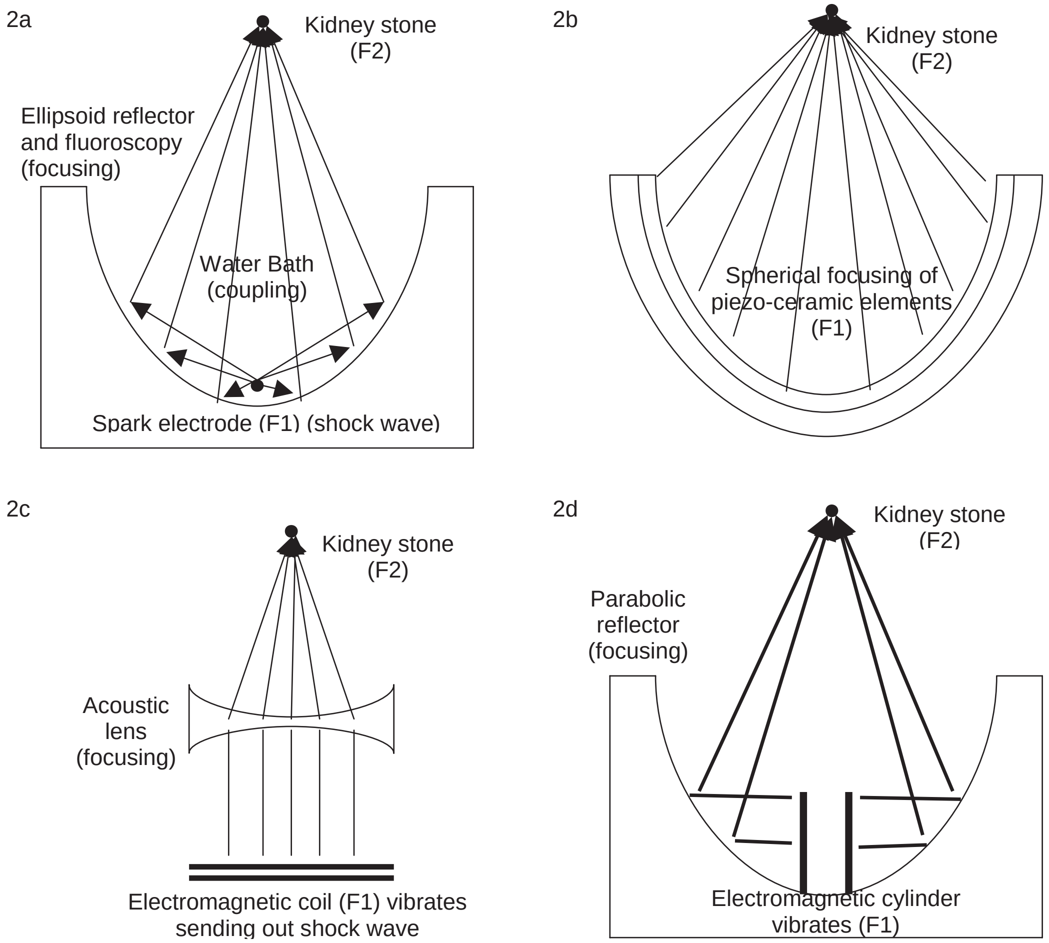

2a
Kidney stone (F2)
Ellipsoid reflector and fluoroscopy (focusing)
Water Bath (coupling)
Spark electrode (F1) (shock wave)
2b
Kidney stone (F2)
Spherical focusing of piezo-ceramic elements (F1)
2c
Kidney stone (F2)
Acoustic lens (focusing)
Electromagnetic coil (F1) vibrates sending out shock wave
2d
Kidney stone (F2)
Parabolic reflector (focusing)
Electromagnetic cylinder vibrates (F1)

Figure 2. Schematic of different shock wave lithotripters. (2a) Electrohydraulic lithotripsy. Spark electrode generates shock wave which is focused at F2 by an ellipsoid reflector. In the original electrohydraulic lithotripter, a water bath served as a coupling medium to allow passage of the shock wave from the source (F1) to the patient and stone (F2). Fluoroscopy or ultrasound can be used to focus to identify and focus the shock waves on the stone. 2b) Piezoelectric lithotripsy. Many ceramic elements are arranged in a spherical pattern. When energy passes through them, they vibrate and send shock waves through a coupling medium to the stone. 2c and d) Electromagnetic lithotripsy. An electromagnetic coil (2c) or cylinder (2d) are stimulated to vibrate by passage of electric current. Their shock waves are focused on the stone (F2) by an acoustic lens (2c) or a parabolic reflector.

device to focus the shock wave at a focal point, a coupling medium, and a stone localization system. (Fig. 2a) The original electrohydraulic lithotripter utilized a spark plug energy generator with an elliptical reflector for focusing the shock waves. A water bath transmitted the shock waves to the patient with stone localization provided by biplanar fluoroscopy. Modification of the four basic components of this first generation lithotripter has provided the development of second and third generation devices that are currently available.

Shock Wave Generation and Focusing

While all lithotripters share the four aforementioned features, it is the mode of shock wave generation that determines the actual physical characteristics of that particular device. The types of energy sources differ on how efficient they are in fragmenting stones and how many treatments are required to adequately treat the stone. Figure 2 diagrammatically summarizes the three different shock wave generation sources used in commercially available lithotripters.

Electrohydraulic Generator. In the original Dornier HM3 lithotripter, the electrohydraulic generator (Fig. 2a) was located at the base of a water bath and produced shock waves by electric spark discharges of 15,000–25,000 V of 1 μs duration. This high voltage spark discharge produced the rapid evaporation of water that created a shock wave by expanding the surrounding fluid. The generator was located in an ellipsoidal reflector that concentrated the reflected shock waves at a second focal point, F2, with F1 being the point of origin of the primary shock waves. While the HM3 continues to be the gold standard lithotripter for stone fragmentation, the short half-life of its electrode results in variable pressure between shocks the longer it is used. In addition, minimal displacement of the electrode, as it deteriorates at F1, results in displacement of the F2 resulting in inaccurate focusing of the shock wave on the stone. The need for frequent replacement of the electrode increases the cost of electrohydraulic lithotripsy.

Piezoelectric Generator. Piezoelectric shock waves are generated (Fig. 2b) by the sudden expansion of ceramic elements excited by a high frequency, high voltage pulse. Thousands of these elements are placed along the inner surface of a hemisphere at the base of a pool of water. While each of these elements moves only slightly in response to a pulse of electrical energy, the summation of the simultaneous expansion of multiple elements results in a high energy shock wave directed to the focal area, at the center of the sphere. The shock wave is propagated through either a small water basin or a water-filled bag to the focal point, F2. The spherical focusing mechanism of the piezoelectric lithotripters provides a wide area of shock wave entry at the skin surface, which causes minimal patient discomfort, but a very small focal area with the smallest amount of energy at F2 compared to other energy sources (7). The small focal area necessitates greater precision to focus and fragment the stone inside the kidney.

Electromagnetic Generator. In electromagnetic devices (Fig. 2c and d), shock waves are generated when an electrical impulse moves a thin, spherical metallic membrane, which is housed within a cylindrical shock tube. The resulting shock wave, produced in the water-filled shock tube, passes through an acoustic lens and is thereby directed to the focal point, F1 (Fig. 2c). The shock wave is coupled to the body surface with a moveable water cushion and coupling gel (8). Alternatively, when energy is passed through a cylindrical coil, the resulting magnetic field pushes away the surrounding cylindrical membrane producing a shock wave that can be focused by a parabolic reflector (Fig. 2d). While these devices produce reliable shock waves with consistent pressures and focus on F2, they also produce small focal regions that may result in reduced stone fragmentation and higher tissue parenchymal injury.

Shock Wave Focusing. Shock waves must be focused in order to concentrate their energy on a calculus. The type of shock wave generation dictates the method of focusing used. Machines that utilize point sources, such as spark-gap electrodes (electrohydraulic lithotripters), generate shock waves that travels in an expanding circular pattern. All of these machines use ellipsoid reflectors for focusing shock waves at the second focal point, F2.

Since a single piezoelement produces a very small amount of energy, larger transducers with multiple ceramic elements are required for piezoelectric lithotripters. The array of elements is positioned in a spherical dish that allows focusing in a very small focal region, F1. Finally, the vibrating metal membranes of the electromechanical lithotripters produce an acoustical plane wave that uses an acoustic lens for focusing the shock wave at F1.

Coupling Medium

The original Dornier HM3 machine utilized a 1000 L water bath to transmit shock waves into the patient. This method of patient coupling required unique positioning of the patient, since the anesthetized subject had to be lowered into the tub and the calculus accurately positioned at the second focal point. Second generation lithotripters were designed to alleviate the physiologic, functional, and economic problems of a large water bath. Current models utilize an enclosed water cushion, or a totally contained shock tube, to allow simplified positioning and “dry” lithotripsy (9).

Stone Localization

Stone localization during lithotripsy is accomplished with either fluoroscopy or ultrasonography. Fluoroscopy provides the urologist with a familiar modality and has the added benefits of effective ureteral stone localization. However, fluoroscopy requires more space, carries the inherent risk of ionizing radiation to both the patient and medical staff and is not useful in localizing radiolucent calculi (certain noncalcium-containing stones are not seen on fluoroscopy or conventional radiographs).

Ultrasonography utilizes sound waves to locate stones. These sound waves are generated at a source. When they encounter different tissue densities, part of the sound wave is reflected back and these reflected waves are used to generate a two dimensional image that can be used to focus the shock wave on a calculus. Sonography-based lithotripters offer the advantages of stone localization with continuous monitoring and effective identification of radiolucent stones, without radiation exposure (7). Additionally, ultrasound has been documented to be effective in localizing stone fragments as small as 2–3 mm, and is as good or better than routine radiographs to assess patients for residual stone fragments following lithotripsy (8). The major disadvantages of ultrasound stone localization include the basic mastery of ultrasonic techniques by the urologist and the inherent difficulty in localizing ureteral stones. While there are several systems that utilize both ultrasound and fluoroscopy to aid in stone localization, many commercially available lithotripters now use a modular design in which the fluoroscopy unit is not attached to the lithotripter reducing storage space as well as allowing use of the fluoroscopy unit for other procedures.

MECHANISMS OF STONE FRAGMENTATION

Due to recent advances made in the basic science of shock wave lithotripsy, there is a better understanding of how shock waves result in stone fragmentation. Both the positive and negative portions of the shock wave (Fig. 1) are critical in stone fragmentation and also play a role in renal tissue injury. The four mechanisms described for stone comminution include compressive fracture, spallation, cavitation, and dynamic fatigue. As the calculus develops *in vivo*, it is formed by both crystallization of minerals as well as organic matrix material. This combination forms an inhomogeneous and imperfect material that has natural defects. When the shock wave encounters this inhomogeneous structure, the force generated in the plane of the shock wave places stress on these imperfections resulting in compression-induced tensile cracking. This mechanism is known as compressive fracture. Spallation, another mechanism of shock wave induced stone fragmentation, occurs when the shock wave encounters fluid behind the stone and part of the wave is reflected back onto the stone placing tensile stress on the same imperfections.

Shock waves cause bubbles to form on the surface of the stone. These bubbles grow during the negative or tensile component of the shock wave. When the positive component of the next wave passes, these bubbles violently collapse releasing their energy against the stone surface as secondary shock waves and/or microjets. This phenomenon, known as cavitation, represents the third mechanism of stone fragmentation. Dynamic fatigue describes the sum of accumulated damage to the stone that coalesce to result in stone fragmentation and eventually destruction of the stone (10,11).

CLINICAL RESULTS OF SHOCK WAVE LITHOTRIPSY

Because many experts continue to consider the Dornier HM3 (the original electrohydraulic lithotripter, Dornier MedTech America, Inc., Kennesaw, GA) the gold standard in lithotripsy, the available clinical literature comparing the Dornier HM3 (first generation) lithotripter to other commercially available lithotripters is summarized in Table 1. This type of summary does not truly allow a comparison between different lithotripters currently in use. Yet, this comparison demonstrates that modifications to second and third generation lithotripters have traded better patient comfort for a lessening of stone free rates. This current clinical data has been one of the driving forces behind the modifications that are currently being investigated to improve SWL.

Summarized results of five large early series on the clinical efficacy of SWL using the unmodified Dornier HM3 lithotripter demonstrate stone free rates for renal pelvis (RP), upper calyx (UC), middle calyx (MC), and lower calyx (LC) stones were 76% (48–85), 69% (46–82), 68% (52–76), and 59% (42–73), respectively. Stone free rates in these series were better for smaller stones (0–10 mm) and RP stones with relatively poorer stone free rates for LC stones. In comparison, results of five large published series on the Siemens Lithostar, a lower power machine demonstrated stone free rates for RP, UC, MC, and LC stones of 69% (55–80), 67% (46–90), 63% (43–82), and 60% (46–73). A comparison of integrated stone free rates stratified by size in a regression model of the HM3 and Lithostar found significantly greater stone free rates across all stone sizes for the HM3 lithotripter (11). While most studies have evaluated the efficacy of SWL for adults, stone free rates with the HM3 are similar for children (21). The limited number of comparative studies of newer machines and the explosion in the number of commercially available lithotripters makes it difficult to assess their clinical efficacy.

While the HM3 has been shown to produce excellent stone free rates for renal calculi, there continues to be debate on the clinical efficacy of SWL for ureteral stones. The main problem is that stones in the ureter are more difficult to locate and therefore more difficult to target with the shock wave. However, several studies have demonstrated stone free rates close to 100% for the treatment of proximal ureteral stones with SWL (22). However, stone free rates appear to decline to 70% for mid-ureteral stones for many lithotripters (23). Treatment of distal ureteral stones with SWL typically involves a prone or a modified sitting position to allow shock wave targeting of the stone below the pelvic brim. Stone free rates of distal ureteral stones with the HM-3 lithotripter have been as high as 85–96% (24). Endoscopic methods (ureteroscopy) can also be employed for ureteral stones, especially those located in the distal ureter with equivalent or better stone free rates.

While many of the lithotripters sighted in Table 1 are no longer in clinical use or have been updated, these studies clearly demonstrated several key points. The HM3 continues to provide equivalent and likely superior stone free rates when compared to other lithotripters in studies. While most commercially available lithotripters provide acceptable stone free rates for stones located within the kidney, these stones often require more treatment sessions and adjunctive procedures to achieve the stone free rates of the HM3 device. Additionally, success rates with all lithotripters declines the further the stone progresses down the ureter and poses positioning challenges with alternative methods indicated to remove ureteral stones.

TISSUE INJURY WITH CLINICAL LITHOTRIPSY

Clinical experience treating patients with SWL has demonstrated that while SWL is generally safe, shock waves can cause acute and chronic renal injury. This concept has been confirmed by multiple animal studies and a few human clinical studies (11). Originally, shear stress to the tissue was believed to be the main cause of this renal injury. However, recent studies have sugggested that SWL induced renal injury is a vascular event induced by vasoconstriction or cavitation-induced injury to the microvasculature (25). Additionally, this initial tissue damage may promote further renal injury via the effects of free radical production (26). Acute damage to the kidney appears to be mainly a vascular insult.

While clinicians have long recognized the acute effects of SWL, most have believed that there was no long-term sequela to shock wave treatment. The > 25 years of clinical experience with SWL serves as a testament to its safety and effectiveness. However, several chronic problems may develop as a consequence of SWL. Table 2 summarizes the acute and chronic effects of SWL. Perhaps the most serious long-term problem of SWL is the increased risk of hypertension. A review of 14 studies on the impact of SWL on blood pressure demonstrated that when stratified according to the number of shock waves administered, higher doses of shock waves seem to correlate with a greater likelihood of increased rates of new-onset hypertension or changes in diastolic blood pressure (11). The impact of hypertension on cardiovascular disease including the risk of myocardial infarction, stroke, and renovascular disease make this a serious long-term effect that needs further investigation.

Three mechanisms of SWL induced tissue injury have been reported: cavitation, vasoconstriction, and free radical induced tissue injury. Investigators have demonstrated *in vitro* that cavitation bubble expansion can rupture artificial blood vessels (5). Other investigators have shown in animal models that cavitation takes place in kidneys exposed to shock waves (27). While cavitation bubbles that form on the stone surface contribute to stone fragmentation,

Table 1. Literature Comparison of Lithotripters to Gold Standard Dornier HM3

Study Type	Reference	HM3 Compared to:	Stone Location	Number of Patients	Stone Free Rates (SFR), %	Auxiliary Procedures	Retreatment Rate, %	Comment
Prospective	12	Wolf Piezolith	Kidney	HM3: 334 Wolf: 378	HM3: 75Wolf: 72		HM3: 15.5 Wolf: 45	Wolf required more retreatment, more shocks, treatment rates decreased dramatically for ureteral stones with Wolf
	13	EDAP LT01	Kidney	HM3: 500 EDAP: 500	HM3:77.2-90.4 EDAP: 42.5-87.5		> with EDAP	More sessions, increased shocks required with EDAP
	14	MFL 5000	Kidney	198 total	HM3: 80MFL: 56			Increased subcapsular hematoma and longer treatment times with MFL
	15	Wolf Piezolith 2300	Ureter	70 total	HM3: 74Wolf: 76.6			Comparable 3 month SFR but used plain radiographs for comparison
	16	Siemens Lithostar	Kidney	HM3: 91Siemens:85	HM3: 91Siemens: 65		HM3: 4 Siemens: 13	SFR comparable at 3 months, Increased tissue injury with HM3 by urinary enzymes
Retrospective	17	EDAP LT01 Sonolith 2000	Kidney and ureter	HM3: 70EDAP: 113Sono: 104	HM3: 79EDAP: 82Sono: 79	HM3: 12EDAP: 13Sono: 9	HM3: 4EDAP: 42Sono: 26	SFR at 3 months comparable with HM3 and EDAP, lower for Sonolith, lower retreatment with HM3
	18	Siemens Lithostar, Dornier HM4, Wolf Piezolith 2300, Direx Tripter X-L, Breakstone	Kidney and ureter	Multicenter	comparable between 2nd generation			All were deemed inferior to HM3 in terms of stone free rates
	19	Medstone STS	Kidney	HM3: 5698Med: 8166	HM3: 70Med: 81.5	HM3: 3.1Med: 5.5	HM3: 4.4 Med: 5.2	Slightly better retreatment and need for auxiliary procedures with HM3
	20	Lithotron	Kidney	38 matched pairs	HM3: 79Lithotron: 58	> for Lithotron	> for Lithotron	HM3 superior to Lithotron using matched pair analysis
	11	Siemens Lithostar		Meta-analysis	HM3: 59-76 Siemens: 60-69			Using regression model, SFR better with HM3 across all stone sizes

Table 2. Acute and Chronic Injury with SWL

Acute	Chronic
Renal edema (swelling)	Hypertension(elevated blood pressure)
Hematuria (blood in urine)	Decreased renal function
Subcapsular hematoma	Accelerated stone formation (in animal models)
Decreased renal blood flow Altered renal function: Impaired urine concentration Impaired control of electrolytes	Renal scar formation

their formation in other locations (tissue, blood vessel lumen) is an unwanted end product that results in tissue injury.

Recent investigations have elucidated yet another potential mechanism of renal injury secondary to high energy shock waves. Evidence suggests that SWL exerts an acute change in renal hemodynamics (i.e., vasoconstriction) that occurs away from the volume targeted at F2, as measured by a transient reduction in both glomerular filtration rate (GFR) and renal plasma flow (RPF) (28). Prolonged vasoconstriction can result in tissue ischemia and permanent renal damage.

Vasoconstriction and cavitation both appear to injure the renal microvasculature. However, as the vasoconstriction induced by SWL abates, reperfusion of the injured tissue might also result in further tissue injury by the release of free radicals. These oxidants produced by the normal processes of cellular metabolism and cellular injury cannot be cleared and injure the cell membrane destroying cells. Free radical formation has been demonstrated in animal models (26).

It appears that the entire treated kidney is at risk of renal damage from SWL-induced direct vascular and distant vasoconstrictive injury, both resulting in free radical formation. Although previous studies have suggested that the hemodynamic effects are transient in nature in normally functioning kidneys, patients with baseline renal dysfunction may be at significant risk for permanent renal damage (28). Patients of concern may be pediatric patients, patients undergoing multiple SWL treatments to the same kidney, patients with solitary kidneys, vascular insufficiency, glomerulosclerosis, glomerulonephritis, or renal tubular insult from other causes.

SHOCK WAVE LITHOTRIPSY ADVANCES

Based on a better understanding of cavitation in stone fragmentation as well as the role of cavitation, vasoconstriction, and free radical formation in SWL-induced tissue injury, several groups are investigating ways in which SWL can be clinically more effective and safe. In general, these advancements involve changes to the shock wave itself, by modifying the lithotripter, alterations in treatment technique, improvements in stone fragmentation / passage or the reduction in tissue injury through medical adjuncts, and improved patient selection.

Changes to the Lithotripter

There are two major mechanical modifications that can improve stone comminution, based on our current understanding of acoustic physics. One is to enhance the compressive component of the shock wave. The original HM3 relies on a high energy output and thus the compressive component to achieve stone fragmentation. The downside of this effect, from clinical experience, is patient discomfort and potential renal tissue injury. Alternatively, one can improve stone comminution by altering the tensile component of the shock wave and thus better control cavitation. Below, several ways are describe in which investigators are modifying the shock wave to improve comminution, with decreased renal tissue injury.

Several investigators have modified lithotripters to alter the negative portion of the shock wave that is responsible for cavitational-induced stone fragmentation and tissue injury. In one study, a reflector insert is placed over the original reflector of an electrohydraulic lithotripter to create a second shock wave that arrives behind the original shockwave, thus partially canceling out the negative component of the shock wave. These investigators found that this modification reduced phantom blood vessel rupture, while preserving stone fragmentation *in vitro* (29). Similarly, an acoustic diode (AD) placed over the original reflector, has the same impact as this modified reflector (30).

However, because reducing the tensile component of the shock wave weakens that collapse of bubbles at the stone surface, two groups have designed piezoelectric inserts into an electrohydraulic lithotripter that send small, focused shock waves at the time of bubble collapse near the stone surface thus, intensifying the collapse of the cavitation bubble without injuring surrounding tissue (29).

Another way in which investigators have modified the shock wave is by delivering shock waves from two lithotripters to the same focal point, F2. Dual pulse lithotripsy has been evaluated by several investigators both *in vitro*, animal models and in clinical trials. Several investigators have demonstrated in an *in vitro* model that the cavitation effect became more localized and intense with the use of two reflectors. Also, the volume and rate of stone disintegration increased with the use of the two reflectors, with production of fine (< 2 mm) fragments (31). In both animal models and clinical studies, dual pulse lithotripsy has been shown to improve stone fragmentation with reduced tissue injury.

Modifications to Treatment Strategy

The original Dornier HM3 lithotripter rate was synchronized with the patient's electrocardiogram so that the shock rate did not exceed the physiologically normal heart rate. Experience with newer lithotripters has revealed that ungating the shock wave delivery rate results in few cardiac abnormalities. As a result, there has been a trend to deliver more shock waves in a shorter period of time. However, increasing doses of shock wave energy at a higher rate may have the potential to increase acute and chronic renal damage.

As a result, several investigators have evaluated ways in which the treatment strategy can be modified to optimize SWL treatment. One proposed strategy is altering the rate of shock wave delivery. Several investigators have reported that 60 shocks $\cdot$ min^{-1} at higher intensities resulted in the most efficient stone fragmentation than 120 shocks · $\min^{-1}$. This has been confirmed *in vitro*, in animal models and also in randomized clinical trials. These studies speculate that at increased rates, more cavitation bubbles were formed in both the fluid and tissue surrounding the stone that did not dissipate between shocks. As a result, these bubbles scattered the energy of subsequent shocks resulting in decreased efficiency of stone fragmentation (32).

In order to acclimate patients to shock waves in clinical treatment, lower energy settings are typically used and gradually increased. A study investigating whether increasing voltage (and thus increasing treatment dose) impacted on stone fragmentation has been performed *in vitro* and in animal models. Stones fragmented into smaller pieces when they were exposed to increasing energy compared to decreasing energy. The authors speculate that the low voltage shock waves "primed" the stone for fragmentation at higher voltages (33). In addition, animals exposed to an increasing voltage had less tissue injury than those kidneys exposed to a decreasing or stable dose of energy. While this treatment strategy has not been tested clinically, it might be able to improve *in vivo* stone comminution while decreasing renal parenchymal injury (34).

In the same vein, several studies have reported that pretreating the opposite pole of a kidney with a low voltage dose of shock waves (12 kV), prior to treating a stone in the other pole of a kidney with a normal dosage of shock waves, reduced renal injury when as little as 100 low voltage shocks were delivered to the lower pole. It is believed that the low voltage shock waves causes vasoconstriction which protects the treated pole of the kidney from hemorrhagic injury (35).

Other SWL treatment modifications being tested include aligning the shock wave in front of the stone in order to augment cavitation activity at the stone surface or apply overpressure in order to force cavitation bubble collapse. While these techniques have only been investigated *in vitro*, these alterations in shock wave delivery, as well as the previous treatment strategies demonstrate how an improved understanding of the mechanisms of SWL can enhance stone comminution and potential reduce renal tissue injury (36,37).

Adjuncts to Improve SWL Safety and Efficacy

Antioxidants. A number of studies have investigated the role of antioxidants in protecting against free radical injury to renal parenchyma (38). Table 3 summarizes the results of various *in vitro* and *in vivo* studies on the use of antioxidants to protect against SWL-induced renal injury due to free radicals. While these studies are intriguing, further clinical trials will be needed to evaluate potential antioxidants for use in patients undergoing SWL.

Improving Stone Fragmentation. Another potential way to improve stone fragmentation is to alter the stone's susceptibility to shock wave energy. One group has demonstrated that after medically pretreating stone in an *in vitro* environment, one could achieve improved stone fragmentation. These data suggest that by altering the chemical environment of the fluid surrounding the stones it is possible to increase the fragility of renal calculi *in vitro* (49). Further studies are warranted to see if calculi can be clinically modified, prior to SWL therapy, in the hopes of enhanced stone fragmentation.

Table 3. Investigated Antioxidants Providing Protection against SWL-Induced Free Radical Injury

Reference	Study Type	Antioxidant
39	*In vitro*	nifedipine, verapamil, diltiazem
40	*In vitro*	Vitamin E, citrate
41	*In vitro*, animal	Selenium
42	Animal	Verapamil
43	Animal	Verapamil
26	Animal	Allopurinol
44	Animal	Astragulus membranaceus, verapamil
45	Human	Antioxidant vitamin
46,47	Human	Verapamil, nifedipine
48	Human	Mannitol

Improving Stone Expulsion. Several reports have demonstrated that calcium channel blockers, steroids, and alpha blockers all may improve spontaneous passage of ureteral stones. (ref) Compared to a control group, investigators found improved stone clearance and shorter time to stone free in patients treated with nifedipine or tamsulosin following SWL compared to the control group. Additionally, retreatment rates were lower (31%) for the medical treatment group compared to the control group (51%). While expulsive therapy appears to offer improved outcomes following SWL for ureteral stones, confirmation with a randomized controlled study is needed (50). Another intriguing report from the same group involves the use of Phyllanthus niruri (Uriston) to improve stone clearance of renal stones following SWL. This medication is derived from a plant used in Brazilian folk medicine that has been used to treat nephrolithiasis. Again, stone free rates were improved with the administration of Uriston following SWL compared to a control group with apparently the greatest effect seen in those patients treated with lower pole stones (51).

Improving Stone/Patient Selection for SWL. Another way to enhance the efficacy of SWL is improve patient selection. Advances in computed tomography (CT) have allowed better determination of internal stone architecture (52). As a consequence, a few studies have demonstrated that determining the Hounsefield units (i.e., density unit of material on CT) of renal stones on pretreatment, noncontrasted CT could predict stone free rates of patients treated with SWL (53). Current micro-CT and newer multidetector CT scanners have the potential to identify stone composition based on CT attenuation. Therefore, stone compositions that are SWL resistant, such as calcium oxalate monohydrate or cystine stones, can be identified and those patients can be treated with endoscopic modalities, therebye avoiding additional procedures in these patients (54). Clinical trials utilizing this concept will need to be performed.

Other factors, such as the distance of the stone from the skin, weight of the patient, and other imaging modalities, are being investigated to help determine who is likely to benefit the most from SWL and which patients should be treated initially with other modalities.

CONCLUSIONS

Shock wave lithotripsy has revolutionized the way in which urologists manage urinary calculi. Patients can now be treated with minimal discomfort using an outpatient procedure. While all lithotripters rely on the same fundamental principles of acoustic physics, second and third generation lithotripters appear to have traded patient comfort and operator convenience for reduced stone free rates, as compared to the original HM3 lithotripter. In addition, mounting evidence demonstrates that SWL has both acute and chronic impact on renal function and blood pressure as a result of renal scarring.

Basic science research has provided insight into how SWL results in stone comminution as well as renal tissue injury. While the compressive component of the shock wave causes stone comminution, it is apparent that the tensile component plays a critical role in creating passable stone fragments. These same forces cause tissue injury by damaging the microvasculature of the kidney. This knowledge has resulted in several novel modifications to improve both stone free rates as well as SWL safety. Mechanical modifications to lithotripsy have focused on controlling cavitation. Preventing bubble expansion in blood vessels while intensifying bubble collapse near the stone surface has been demonstrated to achieve improved stone comminution with decreased tissue injury *in vitro* and in animal models. Many of these designs could be adapted to conventional lithotripters. Modification of treatment techniques have also stemmed from our better understanding of SWL. Slowing treatment rates may limit the number of cavitation bubbles that can interfere with the following shock wave. Voltage stepping and alternative-site pretreatment with low dose shock waves, may cause global renal vasoconstriction that prevents cavitational injury to small vessels during treatment. In addition, our understanding that free radicals may be the end culprit in parenchymal damage has suggested that pretreatment with antioxidants may prevent SWL-induced renal injury. Finally, improved CT imaging may allow us to predict which stones and patients are best suited for SWL versus endoscopic stone removal. Further advances will continue to make SWL a major weapon in the war against stone disease for years to come.

BIBLIOGRAPHY

1. Lokhandwalla M, Sturtevant B. Fracture mechanics model of stone comminution in ESWL and implications for tissue damage. Phys Med Biol 2000;45:1923–1940.
2. Coleman AJ, Saunders JE, Crum LA, Dyson M. Acoustic cavitation generated by an extracorporeal shockwave lithotripter. Ultrasound Med Biol 1987;13:69–76.
3. Zhu S, Cocks FH, Preminger GM, Zhong P. The role of stress waves and cavitation in stone comminution in shock wave lithotripsy. Ultrasound Med Biol 2002;28:661–671.
4. Howard D, Sturtevant B. *In vitro* study of the mechanical effects of shock-wave lithotripsy. Ultrasound Med Biol 1997;23:1107–1122.
5. Zhong P, Zhou Y, Zhu S. Dynamics of bubble oscillation in constrained media and mechanisms of vessel rupture in SWL. Ultrasound Med Biol 2001;27:119–134.
6. Chaussy C, et al. First clinical experience with extracorporeally induced destruction of kidney stones by shock waves. J Urol 1982;127:417–420.
7. Preminger GM. Sonographic piezoelectric lithotripsy: More bang for your buck. J Endourol 1989;3:321–327.
8. Abernathy BB, et al. Evaluation of residual stone fragments following lithotripsy: Sonography versus KUB. In: Lingeman JE, Newman DM., editors. Shock Wave Lithotripsy II. New York: Plenum Press; 1989. pp 247–254.
9. Cartledge JJ, Cross WR, Lloyd SN, Joyce AD. The efficacy of a range of contact media as coupling agents in extracorporeal shockwave lithotripsy. BJU Int 2001;88:321–324.
10. Eisenmenger W. The mechanisms of stone fragmentation in ESWL. Ultrasound Med Biol 2001;27:683–693.
11. Lingeman JE, Lifshitz DA, Evan AP. Surgical Management of Urinary Lithiasis. In: Walsh PC, Retik AB, Vaughan ED, Jr., Wein AJ., editors. Campbell's Urology. Philadelphia: Saunders; 2002. p 3361–3451.
12. Rassweiler J, et al. Wolf Piezolith 2200 versus the modified Dornier HM3. Efficacy and range of indications. Eur Urol 1989;16:1–6.
13. Sofras F, et al. Extracorporeal shockwave lithotripsy or extracorporeal piezoelectric lithotripsy? Comparison of costs and results. Br J Urol 1991;68:15–17.
14. Chan SL, et al. A prospective trial comparing the efficacy and complications of the modified Dornier HM3 and MFL 5000 lithotriptors for solitary renal calculi. J Urol 1995;153:1794–1797.
15. Francesca F, et al. Ureteral lithiasis: In situ piezoelectric versus in situ spark gap lithotripsy. A randomized study. Arch Esp Urol 1995;48:760–763.
16. Graber SF, Danuser H, Hochreiter WW, Studer UE. A prospective randomized trial comparing 2 lithotriptors for stone disintegration and induced renal trauma. J Urol 2003;169: 54–57.
17. Tan EC, Tung KH, Foo KT. Comparative studies of extracorporeal shock wave lithotripsy by Dornier HM3, EDAP LT 01 and Sonolith 2000 devices. J Urol 1991;146:294–297.
18. Bierkens AF, et al. Efficacy of second generation lithotriptors: A multicenter comparative study of 2,206 extracorporeal shock wave lithotripsy treatments with the Siemens Lithostar, Dornier HM4, Wolf Piezolith 2300, Direx Tripter X-1 and Breakstone lithotriptors. J Urol 1992;148:1052–1056. Discussion 1056–1057.
19. Cass AS. Comparison of first generation (Dornier HM3) and second generation (Medstone STS) lithotriptors: Treatment results with 13,864 renal and ureteral calculi. J Urol 1995;153:588–592.
20. Portis AJ, et al. Matched pair analysis of shock wave lithotripsy effectiveness for comparison of lithotriptors. J Urol 2003;169:58–62.
21. Cass AS. Comparison of first-generation (Dornier HM3) and second-generation (Medstone STS) lithotripters: Treatment results with 145 renal and ureteral calculi in children. J Endourol 1996;10:493–499.
22. Robert M, A'Ch S, Lanfrey P, Guiter J, Navratil H. Piezoelectric shockwave lithotripsy of urinary calculi: Comparative study of stone depth in kidney and ureter treatments. J Endourol 1999;13:699–703.
23. Marguet CG, Springhart WP, Auge BK, Preminger GM. Advances in the surgical management of nephrolithiasis. Minerva Urol Nefrol 2004;56:33–48.
24. Rodrigues Netto Junior N, Lemos GC, Claro JF. *In situ* extracorporeal shock wave lithotripsy for ureteral calculi. J Urol 1990;144:253–254.
25. Evan AP, et al. Shock wave lithotripsy-induced renal injury. Am J Kidney Dis 1991;17:445–450.
26. Munver R, et al. *In vivo* assessment of free radical activity during shock wave lithotripsy using a microdialysis system: The renoprotective action of allopurinol. J Urol 2002;167:327–334.

27. Evan AP, et al. *In vivo* detection of cavitation in parenchyma of the pig kidney during shock wave lithotripsy. American Urological Association Annual Meeting. Orlando (FL): 2002. p 1500.
28. Willis LR, et al. Effects of SWL on glomerular filtration rate and renal plasma flow in uninephrectomized minipigs. J Endourol 1997;11:27–32.
29. Zhou Y, Cocks FH, Preminger GM, Zhong P. Innovations in shock wave lithotripsy technology: updates in experimental studies. J Urol 2004;172:1892–1898.
30. Zhu S, et al. Reduction of tissue injury in shock-wave lithotripsy by using an acoustic diode. Ultrasound Med Biol 2004; 30:675–682.
31. Sheir KZ, et al. Evaluation of synchronous twin pulse technique for shock wave lithotripsy: determination of optimal parameters for *in vitro* stone fragmentation. J Urol 2003; 170:2190–2194.
32. Paterson RF, et al. Stone fragmentation during shock wave lithotripsy is improved by slowing the shock wave rate: studies with a new animal model. J Urol 2002;168:2211–2215.
33. McAteer JA, et al. Voltage-Stepping During SWL Influences Stone Breakage Independent of Total Energy Delivered: *In vitro* studies with model stones. American Urological Association Annual Meeting. Chicago: 2003. p 1825.
34. Maloney M, et al. Treatment strategy improves the *in vivo* stone comminution efficiency and reduces renal tissue injury during shock wave lithotripsy. American Urological Association Annual Meeting. San Antonio (TX): 2005. p 1108.
35. Willis LR, et al. Same-pole application of low- and high-energy shock waves protects kidney from swl-induced tissue injury. American Urological Association Annual Meeting. San Francisco: 2004. p 1114.
36. Sapozhnikov OA, et al. Effect of overpressure and pulse repetition frequency on cavitation in shock wave lithotripsy. J Acoust Soc Am 2002;112:1183–1195.
37. Sokolov DL, et al. Prefocal alignment improves stone comminution in shockwave lithotripsy. J Endourol 2002;16:709–715.
38. Preminger GM. Review: *in vivo* effects of extracorporeal shock wave lithotripsy: animal studies. J Endourol 1993;7: 375–378.
39. Jan CR, Chen WC, Wu SN, Tseng CJ. Nifedipine, verapamil and diltiazem block shock-wave-induced rises in cytosolic calcium in MDCK cells. Chin J Physiol 1998;41:181–188.
40. Delvecchio F, et al. Citrate and Vitamin E Blunt the SWL Induced Free radical surge in an in-vitro MDCK cell culture model. American Urological Association Annual Meeting. San Francisco: 2004. p 1120.
41. Strohmaier WL, Lahme S, Bichler KH. Amelioration of high energy shock wave induced renal tubular injury by selenium-an in vivo study in rats. American Urological Association Annual Meeting. Anaheim (CA): 2004. p 1529.
42. Yaman O, et al. Protective effect of verapamil on renal tissue during shockwave application in rabbit model. J Endourol 1996;10:329–333.
43. Willis LR, et al. Effects of extracorporeal shock wave lithotripsy to one kidney on bilateral glomerular filtration rate and PAH clearance in minipigs. J Urol 1996;156:1502–1506.
44. Sheng BW, et al. Astragalus membranaceus reduces free radical-mediated injury to renal tubules in rabbits receiving high-energy shock waves. Chin Med J (Engl) 2005;118:43–49.
45. Kehinde EO, et al. The effects of antioxidants on renal damage occuring during treatment of renal calculi by lithotripsy. American Urological Association Annual Meeting. San Antonio (TX): 2005. p 1698.
46. Strohmaier WL, et al. Protective effect of verapamil on shock wave induced renal tubular dysfunction. J Urol 1993;150:27–29.
47. Strohmaier WL, et al. Limitation of shock-wave-induced renal tubular dysfunction by nifedipine. Eur Urol 1994;25:99–104.
48. Ogiste JS, et al. The role of mannitol in alleviating renal injury during extracorporeal shock wave lithotripsy. J Urol 2003;169: 875–877.
49. Heimbach D, et al. The use of chemical treatments for improved comminution of artificial stones. J Urol 2004;171: 1797–1801.
50. Micali S, et al. Efficacy of expulsive medical therapy using nifedipine or tamsulosin after shock wave lithotripsy of ureteral stones. American Urological Association Annual Meeting. San Antonio (TX): 2005. p 1680.
51. Antonio C, et al. May Phyllanthus niruri (Uriston) affect the efficacy of ESWL on renal stnoes? A prospective, randomised short term study. American Urological Association Annual Meeting. San Antonio (TX): 2005. p 1696.
52. Zarse CA, et al. Nondestructive analysis of urinary calculi using micro computed tomography. BMC Urol 2004;4:15.
53. Saw KC, et al. Calcium stone fragility is predicted by helical CT attenuation values. J Endourol 2000;14:471–474.
54. Williams JC, et al. Progress in the use of helical CT for imaging urinary calculi. J Endourol 2004;18:937–941.

See also MINIMALLY INVASIVE SURGERY; ULTRASONIC IMAGING.

LIVER TRANSPLANTATION

PAUL J. GAGLIO
Columbia University College
of Physicians and Surgeons
New York, New York

INTRODUCTION

From a conceptual perspective, liver transplantation involves the replacement of a diseased or injured liver with a new organ. Historically, liver transplantation has emerged from an experimental procedure deemed "heroic" therapy for patients not expected to survive, to the treatment of choice with anticipated excellent long-term outcomes for patients with end stage liver disease. This article will outline the history of and indications for liver transplantation, delineate short- and long-term complications associated with the procedure, and discuss the role of immunosuppressive therapy, intrinsic to the long-term success of the procedure.

HISTORY

Historically, the most significant and persistent impediment to liver transplantation has been the availability of suitable organs. Up until the early 1960s, "death" was defined as cessation of circulation, and thus, donation from deceased donors was thought to be both impractical and impossible, as organs harvested from pulseless, nonperfusing donors would not function when transplanted, due to massive cellular injury. The concept of "brain death" and ability to harvest organs from individuals defined as such first occurred at Massachusetts General Hospital in the early 1960s, when a liver was harvested from a patient whose heart was beating despite central nervous system failure. This seminal event led to the development of a new concept; death was defined when cessation of brain function occurred, rather than the cessation of circulation. Thus, brain dead donors with stable blood pressure and the absence of comorbid disease could serve as potential organ donors. Improvements in the ability to preserve and transport organs dramatically increased organ availability, necessitating a centralized system to facilitate procurement and allocation of organs to individuals waiting for transplantation. This was initially provided by SEOPF (the Southeast Organ Procurement Foundation), founded in 1968, from which UNOS (the United Network for Organ Sharing) arose. At present, UNOS operates the OPTN (Organ Procurement and Transplantation Network), providing a centralized agency that facilitates recovery and transportation of organs for transplantation, and appropriately matches donors and recipient.

LIVER TRANSPLANTATION: INITIAL RESULTS

The first reported liver transplantation occurred in 1955, in the laboratory of Dr. Stuart Welch (1). In a dog model, an "auxiliary" liver was transplanted into the abdominal cavity, leaving the native liver *in situ*. Between 1956 and 1960, various investigators initiated experiments in different animal models whereby "orthotopic" liver transplantation was performed, achieved by removal of the native liver and implantation of a "new" liver in its place, requiring anastamoses of the donor and recipient hepatic vein and artery, bile duct, and portal vein (see Fig. 1). These initial attempts at liver transplantation refined the surgical procedure, however, graft dysfunction and death of the animals occurred quickly, due to ineffective immunosuppression and eventual rejection of the liver mediated by the animal's immune system (2).

The first human liver transplants were performed by Dr. Thomas Starzl in 1963, at the University of Colorado (3). These early attempts at transplantation highlighted the difficulties associated with extensive abdominal surgery in desperately ill patients, and were associated with poor outcomes, largely due to technical difficulties and the inability to effectively prevent rejection. Similar negative experiences at other centers led to a worldwide moratorium on liver transplantation. However, a major breakthrough in the ability to prevent rejection and prolong the survival of the transplanted liver occurred following the availability of Cyclosporine in 1972 (described below). With continued refinement of the surgical techniques required to perform liver transplantation, combined with the ability to minimize organ rejection, posttransplant outcomes improved significantly. From 1963 to 1979, 170 patients underwent liver transplantation at the University of Colorado; 56 survived for 1 year, 25 for 13–22 years, and several remain alive today 30 years after their surgery. Continued improvement in posttransplantation outcomes were achieved, and thus, in 1983, the National Institutes of Health (NIH) established that liver transplantation was no longer considered an "experimental" procedure, rather, as definitive therapy for appropriately selected patients with end-stage liver disease. Additional advances in immunosuppression (reviewed below) including the discovery of polyclonal and monoclonal antibodies to T-cells or their receptors, and other agents such as Tacrolimus, Mycophenolate Mofetil, and Sirolimus have further improved outcomes.

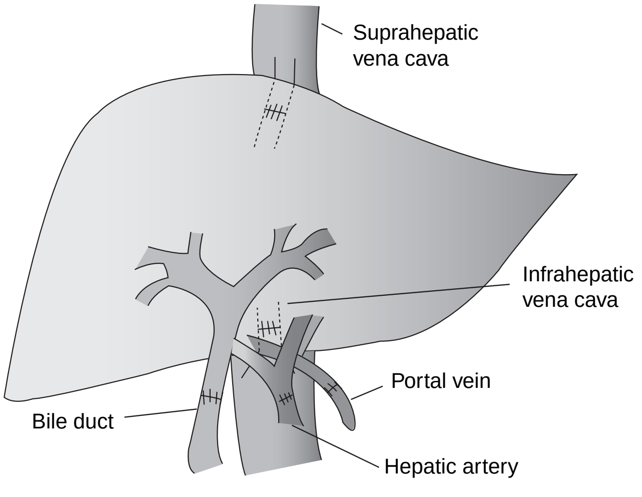


Figure 1. Schematic representation of an orthotopic liver transplant.

INDICATIONS FOR LIVER TRANSPLANTATION

Liver transplantation is an accepted therapeutic modality for complications of chronic liver disease, or acute liver failure. In general, liver transplantation is recommended when a patient with end stage liver disease manifests signs and symptoms of hepatic decompensation, not controlled by alternative therapeutic measures. This is evidenced by

1. Esophageal and/or gastric variceal bleeding, or bleeding from portal hypertensive gastropathy.
2. Hepatic encephalopathy.
3. Spontaneous bacterial peritonitis.
4. Significant ascites.
5. Coagulopathy.

Patients with liver disease complicated by early stage hepatocellular carcinoma (HCC), often defined as either a single lesion <5 cm or not more than three lesions, each <3 cm are also considered candidates for liver transplantation irrespective of evidence of concomitant hepatic decompensation (4).

If a patient meets these initial criteria, further requirements must be realized. It is generally accepted that liver transplantation is indicated if the patient is not moribund and the transplant is likely to prolong life with a $>50\%$ chance of 5-year survival. Furthermore, it is anticipated that the transplant will restore the patient to a range of physical and social function suitable for the activities of daily living. Patients who are suitable candidates should not have comorbid disease with involvement of another major organ system, which would preclude surgery or indicate a poor potential for rehabilitation.

Transplant candidates undergo a thorough psychological assessment prior to liver transplantation. Adequate family and social support must be demonstrated to ensure adherence to the difficult long-term medical regimen that will be required posttransplant. In addition, if a history of substance abuse is present, most transplantation programs require that the patient complete at least 6 months of documented rehabilitation with displayed freedom from alcohol and/or drug recidivism. "Psycho-social" assessment is usually performed by several individuals, including a Psychiatrist or Psychologist and an experienced social worker. In addition, living

Table 1. Diseases Associated with Fulminant Hepatic Failure

Viral Infection	
Frequent	Hepatitis A, B, D, E, Hepatitis Non A-G
Rare	Hepatitis C Cytomegalovirus Epstein Barr virus Herpes simplex virus
Metabolic	Acute fatty liver of pregnancy Reye's SX
Toxin, Drugs	Acetaminophen Nsaid's CCL4 Isoniazid Sodium valproate Methyl DOPA Tetracycline Halothane Amanita phalloides (mushroom poisoning) Yellow Phosphorus "Herbal Medication"
Drug Combos	Acetaminophen and ETOH Acetaminophen and barbiturates, isoniazid trimethoprim, and sulamethoxazole amoxicillin and clavulinic acid
Ischemic	Hepatic artery thrombosis Budd-Chiari Syndrome Right ventricular failure, cardiac tamponade shock
Miscellaneous	Hyperthermia Hellp SX

donor liver transplantation (LDLT), discussed in greater detail below, requires a detailed psychosocial assessment of both recipient and potential donor. In most transplantation centers, an independent donor advocate team consisting of a social worker, internist, and surgeon who are independent of the team evaluating the recipient performs the difficult task of educating a potential donor regarding the risks and benefits of LDLT, assessing motivation to be a donor, and determining if coercion is present.

Another indication for liver transplantation is fulminant hepatic failure, defined as hepatic encephalopathy (confusion) arising in the setting of massive liver injury in a patient without preexisting liver disease. This condition is rapidly fatal unless recovery of hepatic function occurs spontaneously, and thus, emergent liver transplantation may be required. Conditions associated with fulminant hepatic are listed in Table 1.

ETIOLOGY OF LIVER DISEASES REQUIRING LIVER TRANSPLANTATION

Diseases associated with hepatic dysfunction in adults and children are outlined in Tables 2 and 3, respectively. In general, any disease process in adults or children that induces either acute or chronic hepatocellular, biliary, or vascular injury may necessitate liver transplantation. The indications for liver transplantation in children are identical to those in adults, that is, liver transplantation is indicated in the presence of progressive liver disease in patients who fail medical management.

Table 2. Indications for Liver Transplantation

Diseases Effecting Hepatic Parenchyma
Viral hepatitis with cirrhosis (Hepatitis B with or without Delta Virus, Hepatitis C, Non A-E hepatitis)
Autoimmune hepatitis
Alcoholic cirrhosis
Metabolic disorders (Wilson's disease, hemochromatosis, alpha 1 Antitrypsin, Tyrosinemia, protoporphyria, Cystic fibrosis, familial amyloidosis, Neiman-Pick disease)
Fulminant hepatic failure due to any cause
Drug induced liver disease
Diseases Effecting Biliary System
Primary and secondary biliary cirrhosis
Sclerosing cholangitis
Caroli's disease
Relapsing cholangiohepatitis
Choledochal cysts with obstruction and bilary cirrhosis
Hepatic Neoplasia/Malignancies
Patients with nonmetastatic primary hepatocellular carcinoma, with;
A single tumor not > 5 cm
No more than three lesions with the largest lesion < 3 cm
No thrombosis of the portal or hepatic vein
Hemangioendothelioma (confined to the liver)
Neuro endocrine tumors with hepatic involvement
Large hepatic Hemangioma
Miscellaneous Causes
Hepatic vein thrombosis (Budd-Chiari syndrome)
Portal vein thrombosis
Hepatic artery thrombosis
Trauma

Many of the disease processes in adults that induce liver failure are recapitulated in children. However, specific disease states seen in children including metabolic diseases and congenital biliary anomalies represent additional indications for liver transplant. Moreover, liver transplantation is indicated in infants and children if the transplant will prevent or attenuate derangements in cognition, growth, and nutrition. Therefore, children should be considered for liver transplantation when there is evidence that hepatic decompensation is either unavoidable (based on knowledge of the history of the disease itself), imminent, or has already occurred. The clinical scenarios that determine when liver transplantation is required in children can include one or more of the following:

1. Intractable cholestasis.
2. Portal hypertension with or without variceal bleeding.
3. Multiple episodes of ascending cholangitis.
4. Failure of synthetic function (coagulopathy, low serum albumin, low cholesterol).
5. Failure to thrive or achieve normal growth, and/or the presence of cognitive impairment due to metabolic derangements, and malnutrition.
6. Intractable ascites.
7. Encephalopathy.
8. Unacceptable quality of life including failure to be able to attend school, intractable pruritis.
9. Metabolic defects for which liver transplantation will reverse life-threatening illness and/or prevent irreversible central nervous system damage.

Table 3. Additional Indications for Liver Transplantation in Infants and Children

Cholestatic Liver Disease
Obstructive: Biliary Atresia (most common indication for liver transplantation in children)
Intrahepatic: Alagille's Syndrome, Bylers disease, familial cholestatic symptoms
Other
Congenital hepatic fibrosis
Metabolic Diseases

Disease	Defect	Inheritance	Comments
Alpha 1 antitrypsin	Decreased serum A1AT	Codominant	May reverse both liver and lung disease
Wilsons's Disease	Decreased Ceruloplasmin	Autosomal Recessive (AR)	
Tyrosinemia	Fumarylacetoacetate hyrolase	AR	Transplant in fulminant liver failure, or to prevent hepatic neoplasia
Urea cycle defects	Example: ornithine transcarbamylase Arginosuccinate synthetase	x-linked dominant AR	Prevent CNS injury
Galactosemia	Galactose phosphate uridyl transferase	AR	Prevent development of cirrhosis and Hepatoma
Glycogen storage Diseases	Glucose 6 phosphatase	AR	Consider transplant if dietary management not successful
Type 1A	Brancher enzyme		
Type IV			
Familial hypercholesteroloemia	Type 2 A-LDL receptor deficiency	AR	Avoids ASHD
Gaucher's Disease	Glucocerebrosidase	AR	May need combined liver/bone marrow t-plant
Nieman-Pick disease	Sphingomyelinase	AR	
Crigler-Najjar type 1	Uridine diphosphate glucoronly transferase	AR	prevents fatal Kernicterus
Cystic fibrosis	Chloride ion transfer gene	AR	May need combined liver/lung transplant
Hyperoxaluria type 1	Alanine glyoxalate aminotransferase	AR	Usually requires combined liver/kidney
Neonatal Fe storage	Unknown	Varies	Transplant as infant
Hemophilia A and B	Factor VIII/IX	x-linked	Transplant indication varies (?iron overload, factor inhibitor present)
Disorders of bile acid synthesis (Bylers disease)	Unknown	Varies	Transplant indicated if associated with end stage liver disease

10. Life threatening complications of stable liver disease (e.g., hepatopulmonary syndrome).

CONTRAINDICATIONS TO LIVER TRANSPLANTATION (ADULTS AND CHILDREN)

At present, "absolute exclusion" criteria for liver transplantation are evolving. In general, patients with advanced cardiac or pulmonary disease, severe pulmonary hypertension, active substance abuse, coma with evidence of irreversible central nervous system injury, sepsis, or uncorrectable congenital abnormalities that are severe and life threatening are not transplant candidates. In addition, individuals with evidence of extrahepatic malignancy do not meet criteria for transplantation, unless the patient meets standard oncologic criteria for "cure". "Relative" exclusion criteria include renal insufficiency when renal transplantation is not feasible, prolonged respiratory failure requiring $>50\%$ oxygen, advanced malnutrition, primary biliary malignancy, inability to understand the risk/benefits of the procedure, and inability to comply with medications and conform to follow-up regimens. Recent data indicates successful outcomes in HIV infected patients who undergo liver transplantation, a population formerly considered noncandidates for the procedure. However, initial enthusiasm regarding successful transplantation outcomes must be restrained by evidence that HCV recurrence in HIV–HCV coinfected patients may be problematic (5).

RECIPIENT CHARACTERISTICS AND PRIORITIZATION FOR TRANSPLANTATION

Given the relatively stable number of available donor organs in the setting of a rapidly expanding pool of potential recipients, the timing of transplantation is critical. Liver transplantation in a stable patient who is anticipated to do well for many years while waiting for an available organ may not be appropriate, while liver transplantation in a moribund patient with a low probability of posttransplantation survival is similarly inappropriate. Prior to 1997, prioritization for liver transplantation was based on the location where patients received their care (i.e., home, hospital, intensive care unit) and waiting time on the transplant list. In 2002, several policies were instituted by UNOS in an attempt to produce a more equitable organ allocation scheme. Waiting time and whether the patient was hospitalized were eliminated as determinants of prioritization of organ allocation. The "MELD" score (Model for End Stage Liver Disease) a logarithmic numerical score based on the candidate's renal function (creatinine), total bilirubin, and INR (international normalized ratio for prothrombin time) has been shown to be the best predictor of mortality among cirrhotic patients, including those on the transplant waiting list. It was therefore adopted by UNOS as a mechanism to prioritize waiting list candidates. MELD had been validated as a predictor of 3-month survival in diverse groups of patients with various etiologies and manifestations of liver disease (6). Presently, a patient's position on the liver transplantation waiting list is now determined by their MELD score; patients with highest MELD scores are ranked highest on the list. Prospective analysis of the impact of MELD indicates improvement in both the rate of transplantation, pretransplantation mortality, and short-term posttransplantation mortality rates (7). However, retrospective analysis has suggested that posttransplantation survival may be reduced in patients with very high pretransplantation MELD score, particularly in Hepatitis C infected patients (8). Conversely, MELD score effectively delineates when a patient is "too well" for transplantation. A recent review indicates that posttransplantation survival in patients transplanted with a MELD score of <15 is lower than a nontransplanted cohort with similar MELD score (9). Thus, it is clear that careful recipient selection, with attention to pressor and ventilatory requirements, need for dialysis, age, and MELD score are important factors in selecting appropriate candidates for liver transplantation.

LIVER TRANSPLANTATION: SOURCE OF ORGANS

At present, there are three potential types of organ donors specific to liver transplantation, identified as deceased, living, or non-heart beating. Deceased donors (DD) comprise the majority of liver donors. Either by self-identification while living, or after discussion with "next of kin" when donor brain death has been declared, individuals are acknowledged as potential organ donors. Recent data from UNOS indicate that 1- and 3-year patient survival in recipients of DD liver transplant is 81 and 71%, respectively (10). However, despite efforts to maximize utilization of organs acquired from DD including the use of older donors, steatotic (fatty) livers, and livers infected with Hepatitis C or B, a growing disparity exists between the number of available livers and the number of individuals waiting for transplantation. This critical shortage of organs has resulted in both an increase in the waiting time for liver transplantation and death rate among patients on the waiting list. In response, the modalities of adult-to-child and adult-to-adult LDLT have emerged as alternatives to deceased donor liver transplantation (11,12). Adult-to-child LDLT usually involves the removal of the left lateral segment of the liver (~20% of hepatic mass) from an adult donor for implantation into a child, while adult to adult living donor liver transplantation requires that the larger, right lobe of the liver (which accounts for ~50–60% of the hepatic mass) be removed from the donor to ensure adequate hepatic mass in the recipient. Rapid regeneration of the liver remnant in the donor and the partial allograft transplanted into the recipient occurs, to the extent that appropriate liver volume is restored within 1–2 months in both donor and recipient following surgery. Since most pediatric LDLT recipients are <2-years old, they receive a liver graft of adequate or even excessive size, and thus liver insufficiency due to the receipt of inadequate liver mass is rare. In contradistinction, as the recipient of an adult-to-adult living donor liver transplantation receives a graft that must over time grow to an appropriate volume, selection of recipients best able to tolerate transplantation of a "partial" graft is necessary. In appropriately selected pediatric and adult recipients, 1- and 3-year graft and patient survival in individuals who undergo LDLT is similar or superior to DD (10). However, when comparing postoperative complications in recipients of DD versus LDLT, recipients of LDLT have a greater rate of biliary complications including bile leaks and biliary strictures, which occur in 15–32% of patients (13). In addition, the "small-for-size syndrome" manifested as prolonged posttransplantation cholestasis with or without portal hypertension may occur following LDLT, if the graft is of inadequate size (14). Fortunately, the majority of patients who experience this syndrome recover without the requirement of retransplantation.

Recently, significant interest in the utilization of "non-heart beating" donors (donation after cardiac death, DACD) as a potential modality to further increase the pool of available organs has emerged. In contrast to DD who are declared brain dead, DACD are critically ill patients who are not brain dead, but have no expectation of recovery and who based on their own prior wishes or families request are removed from life support. Following cardiac arrest and declaration of death, organs are harvested. There are two types of DACD, "controlled" and "uncontrolled". In the controlled DACD (Maastricht category 3 "death anticipated") the patient is removed from life support and death occurs in the operating room. Once death has been declared, organs deemed suitable for transplantation are rapidly perfused with preservation solution and removed surgically. The uncontrolled DACD (Maastricht category 1 and 2 "death not anticipated") is declared dead after cardiac arrest, rushed to the operating room, and organs are harvested. Uncontrolled DACD are usually not utilized for liver transplantation due to the high rate of primary nonfunction (defined below), usually due to prolonged ischemia of the graft. When utilizing controlled DACD for transplantation, emerging data indicates that recipient and graft survival are diminished when compared to deceased and living donor liver transplantation with a higher incidence of primary nonfunction, biliary injury, and requirement for retransplantation. However, several centers have reported acceptable outcomes when utilizing controlled DACD organs, particularly those without significant ischemia in well-selected recipients (15).

Finally, "domino" transplantation is an option for patients afflicted with familial amyloidotic polyneuropathy (FAP). Familial amyloidotic polyneuropathy is a fatal disease caused by an abnormal amyloidogenic transthyretin (TTR) variant generated by the liver. Liver transplantation in these patients removes the source of the variant TTR molecule, and represents the only known curative treatment. As no intrinsic liver disease exists in patients affected by FAP, the liver explanted from a patient with FAP may be transplanted into another patient, thus, allowing "domino" transplantation. Survival in both recipients of FAP livers and transplanted FAP patients has been reported to be excellent and comparable to survival with OLT performed for other chronic liver disorders (16).

POSTTRANSPLANTATION MANAGEMENT

The complex nature of the surgical procedure utilized to both explant (remove) the diseased, cirrhotic liver and implant (transplant) the new allograft into the recipient make it intuitive that the majority of the early complications following liver transplantation are technical and related to the surgical procedure itself. However, following the first postoperative days, and as patients progress to the first month posttransplantation and beyond, the nature and variety of complications change. Early complications (within the first 2 months) and late complications (beyond 2 months) may negatively affect patient and graft survival (Table 4). Complications specific to the surgical procedure and those that directly affect the transplanted organ are discussed below.

EARLY COMPLICATIONS

Primary Nonfunction and Early Graft Dysfunction

A major threat to the newly transplanted liver is primary graft nonfunction (PNF). This syndrome defined as acidosis, rising INR, progressive elevation in liver transaminases and creatinine, and decreases in mentation occurs when the newly transplanted liver allograft fails to function normally. The mechanisms responsible for this phenomenon are complex, and relate to donor factors, inadequate preservation of the liver, prolonged ischemia, extensive steatosis of the graft, hepatic artery thrombosis (see below) or immune response to the implanted organ (17). In the setting of PNF, a rapid assessment of hepatic artery flow needs to occur, as immediate surgical repair of a thrombosed hepatic artery may reverse PNF. In the absence of hepatic artery thrombosis, emergent retransplantation is required for PNF.

Table 4. Early and Late Complications Following Liver Transplantation

Early
Graft Specific
Primary nonfunction
Early graft dysfunction
Hepatic artery thrombosis
Hepatic and portal vein thrombosis
Preservation injuryBiliary complications: bile leak, biliary stenosis
Acute cellular rejection
Other
Bacterial and fungal infection
CMV infection
Recurrent Hepatitis B and C
Late
Graft Specific
Chronic rejection
Recurrence of primary disease
Other
Hypertension
Hyperlipidemia
Diabetes
Obesity
Cardiac disease
Renal dysfunction
Fungal infection (Cryptococcus, Aspergillus)
CMV
Posttransplant lymphoproliferative disorder
Nonhepatic malignancy: i.e., skin cancer

In contrast to PNF, early graft dysfunction (EGD) is manifested by an early rise in serum transaminases to values > 2000–3000 IU/L, cholestasis with rising Bilirubin levels, without marked coagulopathy or impairment in mental status and renal function. EGD may occur in the setting of ischemic injury or steatosis in the graft, and typically occurs within the first 24–48 h after the transplant. Unlike PNF, the manifestations of EGD usually improve with supportive care, and emergency retransplantation is not necessary.

Hepatic Artery Thrombosis

A potentially devastating posttransplantation complication is hepatic artery thrombosis (HAT). Hepatic artery thrombosis occurs more commonly in pediatric transplant recipients compared to adults due to the technical difficulties associated with the anastomosis of smaller size vessels. In HAT, the immediate postoperative period may be associated with graft failure, elevation in serum liver transaminases, bile leak, hepatic necrosis, and sepsis. Since the blood supply to the biliary tree in the early posttransplant period is principally from the hepatic artery, HAT is frequently associated with irreversible injury to the biliary tract (18). Thus, HAT in the first 7 days after liver transplantation is an indication for emergent artery repair or retransplantation.

Due to the potentially devastating consequences of HAT, most transplant centers screen for this complication with duplex-ultrasound (US) in the immediate posttransplant period. If duplex-US suggests HAT, angiography is usually performed to confirm the diagnosis, and if present, surgical revision of the hepatic artery is required. If surgical repair cannot be achieved, liver retransplantation may be necessary.

PORTAL AND HEPATIC VEIN THROMBOSIS

Though less common than HAT, thrombosis of the portal and/or hepatic veins in the immediate posttransplant period can also adversely affect patient and graft survival. Acute “Budd-Chiari” syndrome due to hepatic vein or vena cava thrombosis is associated with abdominal pain, peripheral edema, and the threat of graft failure, as hepatic congestion in the newly transplanted liver is poorly tolerated. In this circumstance, emergency thrombectomy and surgical revision is required. Acute portal vein occlusion may be associated with exacerbation of preexisting portal hypertension, associated with gastrointestinal bleeding from porto-systemic collateral vessels such as esophageal and gastric varices. Acute portal vein thrombosis is managed by surgical repair, while chronic portal vein thrombosis may be well tolerated. A potential alternative to surgical repair for both hepatic and portal vein stenosis or occlusion is thrombolysis and/or the placement of endovascular stents by an experienced interventional radiologist (19).

ACUTE CELLULAR REJECTION

Rejection of any transplanted organ is a constant threat, as immunologic recognition of the graft as “foreign” may be associated with injury. However, compared to other organs, liver allografts are relatively privileged immunologically, and thus, the incidence and consequences of acute cellular rejection (ACR) are diminished when compared to other solid organs utilized for transplantation. The reported incidence of ACR within the first posttransplant year is 30–50%, in most cases, usually occurring within the first 2–3 weeks postoperatively. The clinical presentation is variable; ACR may be asymptomatic, or associated with fever or abdominal pain. Laboratory findings include elevation or failure of normalization of serum transaminases, usually in association with a rising alkaline phosphatase and/or bilirubin. The diagnosis of acute liver graft rejection is confirmed by liver biopsy and examination of liver histology (20). Conventional histologic criteria associated with ACR include the presence of periportal lymphocytic infiltrate, as well as bile duct and hepatic vascular endothelial cell injury. Most cases of ACR respond to treatment with intravenous bolus glucocorticoids. Approximately 10% of patients with ACR will not improve with intravenous glucocorticoids, requiring the administration of monoclonal or polyclonal anti-T cell antibodies (reviewed below). Mild and moderate ACR may also respond to either increasing the dose of the primary immunosuppressive agent, or switching to an alternate calcineurin inhibitor. This approach has been used with increasing frequency, particularly in patients transplanted for HCV and HBV due to concerns regarding the negative impact of over-immunosuppression on viral recurrence.

BILIARY COMPLICATIONS

Bile leaks and strictures generally occur at the anastomosis of the donor and recipient bile ducts, recognized by a rise in serum bilirubin and/or alkaline phosphatase or by the presence of bile in surgical drains in the immediate posttransplantation period. The incidence of biliary complications is between 5 and 15% following deceased donor liver transplantation. However, between 15 and 30% of patients who undergo living donor liver transplantation develop biliary complications, due to the complexity of the biliary reconstruction required (21). In both deceased and living donor recipients, the majority of bile leaks resolve spontaneously without the need for reoperation. As previously stated, the biliary tree receives the vast majority of its blood supply from the hepatic artery, and thus, the adequacy of hepatic artery blood flow needs to be evaluated in the setting of any biliary injury. If spontaneous resolution of the bile leak does not occur, endoscopic or radiologic placement of a biliary stent across the biliary anastamoses is often successful (22). In some cases, surgical exploration and revision of the biliary anastamoses with a Roux-en-Y choledochojejunostomy may be required.

Anastamotic biliary strictures require careful attention, as if left untreated, cholangitis, graft dysfunction, and eventually secondary biliary cirrhosis may occur. Techniques for management include dilatation and stenting via biliary endoscopy or percutaneous transhepatic cholangiogram by an interventional radiologist. If these modalities are unsuccessful, surgical revision of the biliary anastamosis with a Roux-en-Y choledochojejunostomy may be required. In rare cases with diffuse stricturing, retransplantation may be necessary.

ISCHEMIC AND PRESERVATION INJURY

The newly transplanted liver is always subjected to some degree of ischemic injury (23). Cold (or hypothermic) ischemia is unavoidable, as it occurs prior to transplantation while the liver is cooled in preservation solution, awaiting implantation. Warm (normothermic) ischemia occurs during the transplantation procedure itself, when hepatic blood flow is interrupted to minimize blood loss during transplantation, or when the formerly “cooled” liver is subjected to body temperature during transplantation. Cold ischemia is usually well tolerated, while in contrast, warm ischemia often leads to death of hepatocytes, with resultant elevation in serum transaminases, apoptosis and centrilobular necrosis. In the setting of significant warm ischemia, graft failure may result. Several investigators have noted improvement in ischemic injury and enhanced graft and patient outcomes by employing a technique described as “ischemic preconditioning” defined as a brief period of controlled ischemia followed by a short interval of reperfusion before the actual surgical procedure (24). This is accomplished during liver transplantation by transiently interrupting hepatic inflow by placing a vascular clamp or a loop around the portal triad (i.e., portal vein, hepatic artery, and bile duct), rendering the whole organ ischemic for 10–15 min, after which the clamp is removed and the liver is reperfused. This technique may be of particular benefit in organs with significant steatosis.

Complications Beyond Two Months

Progress in the surgical techniques required to perform transplantation, the treatment of postoperative complications and prevention of rejection have been associated with significant improvements in short-term morbidity and mortality following transplantation. Coincident with improvements in short-term outcomes has been a rise in long-term complications. These complications, including side effects of chronic immunosuppression, neoplasia, and infections are discussed in detail elsewhere. Long-term complications that affect the transplanted liver are discussed below.

CHRONIC REJECTION

Chronic allograft rejection or “vanishing bile duct syndrome” is rare, but in contradistinction to acute cellular rejection, a much more difficult to treat complication. Diagnostic criteria for chronic rejection include bile duct atrophy affecting the majority of bile ducts, with or without bile duct loss. Arterial and venous injury affecting the large branches of the hepatic artery or portal vein (foamy arteriopathy) may also be present (24). Risk factors for chronic liver rejection include transplantation for primary sclerosing cholangitis, primary biliary cirrhosis, HLA mismatch between donor and recipient, and cytomegalovirus infection. Chronic rejection is usually a harbinger of poor outcomes, often resulting in the requirement for retransplantation; altering immunosuppression is rarely associated with improvement.

Recurrence of Primary Disease Following Liver Transplantation

A major challenge to the liver transplant community is recurrence of the primary disease that caused the patients native liver to fail. Diseases that do not recur following liver transplantation include congenital anatomic anomalies (e.g., biliary atresia, polycystic liver disease, Caroli's disease, Alagilles syndrome, congenital hepatic fibrosis) and metabolic diseases of the liver (e.g., Wilson's disease, alpha 1 antitrypsin deficiency). However, all other causes of liver disease including primary biliary cirrhosis, primary sclerosing cholangitis, autoimmune hepatitis, nonalcoholic fatty liver disease, hemochromatosis and alcohol related liver disease have been reported to recur after liver transplantation. In some cases, recurrent disease may lead to significant liver injury with resultant graft failure (26–30). Disease processes most commonly associated with recurrence include viral hepatitis B (HBV) and C (HCV). The recurrence of HBV is associated with uniformly poor outcomes with graft failure and death. Fortunately, recurrence of HBV after liver transplantation can be prevented by administering hepatitis B immune globulin (HBIG) at the time of transplantation and at regular intervals thereafter, with or without the use of antiviral agents such as Lamivudine and Adefovir. In contradisctinction to HBV, HCV recurrence following liver transplantation remains a significant source of morbidity and mortality, with negative impact on post-transplantation outcomes. In patients with active HCV replication prior to transplantation, reacquisition of viremia following transplantation is universal, and histologic injury due to HCV occurs in up to 90% of patients followed for 5 years (31). Although histologic injury in the allograft due to HCV is exceedingly common, disease progression after the development of hepatitis is variable, with some patients experiencing indolent disease and others rapidly progressing to cirrhosis and liver failure. In patients that develop HCV associated cirrhosis posttransplantation, up to 42% will experience decompensation manifested as ascites, encephalopathy, or hepatic hydrothorax, and $<50\%$ of patients survive > 1 year after the development of decompensation (32). It is important to contrast the natural history of HCV before and after transplant; prospective and retrospective data are emerging which indicate that the progression of HCV following liver transplantation is accelerated when compared to the nonimmunosuppressed pretransplant patient population.

Whether HCV recurrence is more severe in recipients of LDLT than in DD recipients is controversial. Although several recent reports indicate that HCV recurrence may be more problematic in recipients of LDLT when compared to DD (33), particularly the cholestatic variant of HCV (34), other authors have noted no differences in outcomes in HCV infected patients who undergo LDLT when compared to DD (35,36). At present, both the optimal timing for transplant in HCV patients and the therapy for recurrent HCV following liver transplantation are incompletely described. Theoretically, eradication of HCV prior to liver transplantation in patients with decompensated liver disease would be beneficial, although in practice, this strategy has been marred by exacerbation of encephalopathy, infections, and other serious adverse events, particularly in patients treated with high dose Interferon and ribavirin (37). A novel approach including initiating therapy with low dose interferon (including Pegylated interferon preparations) and ribavirin with slow escalation in dose may be associated with improved tolerability and efficacy (38). Following liver transplantation, both preemptive therapy prior to the development of histologic injury and directed therapy after the onset of liver injury have been attempted with varying degrees of success. It is important to note, however, that posttransplantation, tolerability of interferon preparations, and ribavirin is suboptimal. Significant leucopenia and anemia are common, likely due to drug induced bone marrow suppression and renal insufficiency potentiating ribavirin induced hemolysis (39).

Immunosuppressive Medications

A cornerstone to posttransplantation management is the ability to prevent or attenuate immunologic rejection of the transplanted graft, which when left untreated, can be associated with graft failure. From a conceptual standpoint, understanding how recognition of the newly engrafted liver as "foreign" occurs, how to modulate immune mediated injury, and at the same time prevent "overimmunosuppression" are critical to achieve optimal post transplantation outcomes. The various immunosuppressive medications and their mechanism of action currently utilized in liver transplant recipients are listed in Table 5. Unfortunately, all immunosuppressive therapy is associated with undesired effects, with the potential for additive effects when agents are combined. In general, most transplant centers utilize three agents to prevent allograft rejection in the immediate posttransplant period, utilizing a combination of a calcineurin inhibitor such as Cyclosporine (CYA) or Tacrolimus (TAC), a second agent such as Mycophenolate mofetil (MMF) or Azathioprine (AZA), and a glucocorticoid such as Prednisone. As patients achieve adequate liver function and freedom from rejection beyond 6-months posttransplantation, satisfactory immunosuppression can be achieved in many patients with monotherapy, usually with a calcineurin inhibitor, although in patients who are at increased risk of rejection such as those with autoimmune hepatitis, primary biliary cirrhosis, or sclerosing cholangitis, long-term immunosuppression is achieved with a combination of a calcineurin inhibitor with either low dose MMF or Prednisone (40).

Corticosteroids

Corticosteroids achieve their desired immunosuppressive affects by the suppression of leukocyte, macrophage, and cytotoxic T-cell activity, and diminution of the effect of

Table 5. Immunosuppressive Agents

Agent	Mechanism of Action	Side Effects
Antilymphocyte globulin Antithymocyte globulin	Depletes circulating lymphocytes	Flu-like symptoms Anaphylaxsis Lymphoproliferative disorders
OKT3	Depletes circulating T cells	Flu-like symptoms Anaphylaxsis Lymphoproliferative disorders
Basiliximab Daclizumab	IL-2 receptor blockade	Infections Gastrointestinal distress Pulmonary edema and bronchospasm (rare)
Cyclosporine	Inactivates calcineurin, decreases IL2 production, Inhibits T-cell activation	Hypertension Renal insufficiency Neuropathy Hyperlipidemia Gingival hyperplasia HirsutismInsulin resistance
Prednisone	Suppression of leukocyte, macrophage, and cytotoxic T-cell activity Decrease cytokines, prostoglandins, and leukotrienes	Hypertension Dyslipidemia Glucose intolerance Bone abnormalities Peptic ulcers Psychiatric disorders
Azathioprine	Inhibits adenosine and guanine production Inhibits DNA and RNA synthesis in rapidly proliferating T cells	Leukopenia Anemia Thrombocytopenia Pancreatitis
Tacrolimus	Inactivates calcineurin, decreases IL2 production, Inhibits T-cell activation	Hypertension Renal insufficiency Insulin resistance Neuropathy Hyperlipidemia
Mycophenolate mofetil	Inhibits of inosine monophosphate dehydrogenase (IMPDH) Prevents T- and B-cell proliferation	Leukopenia Anemia Thrombocytopenia GI side effects
Sirolimus	inhibiting mTOR (target of Rapamycin) Prevents T-cell replication.	Hepatic artery thrombosis Bone marrow suppression Hyperlipidemia Pneumonitis Inhibits wound healing

cytokines, prostaglandins, and leukotrienes. However, hypertension, dyslipidemia, glucose intolerance, bone loss, peptic ulcers and psychiatric disorders are often associated with therapy. Therefore, a strategy to taper and discontinue glucocorticoids within the first 6 months–1 year following transplantation while maintaining adequate levels of calcineurin inhibitor is employed by many transplant centers. This tactic is often altered in patients who undergo liver transplantation secondary to an immunologic disorder such as autoimmune hepatitis, primary biliary cirrhosis and sclerosing cholangitis due to an enhanced risk of acute cellular rejection. In these patients, either long-term use of corticosteroids with an attempt to minimize doses is advocated, or chronic use of MMF or AZA in combination with a calcineurin inhibitor is required.

T-Cell Depleting Agents

In the past, “induction therapy” with antilymphocyte agents such as antilymphocyte globulin or antithymocyte globulin or monoclonal antibody preparations such as OKT3 was utilized immediately after liver transplantation to rapidly induce an immune suppressed state via the rapid destruction of the host’s T cells. However, due to significant systemic side effects including fevers, allergic reactions, serum sickness, and thrombocytopenia, the use of these agents is now usually reserved for the treatment of glucocorticoid resistant rejection, or less commonly, in patients with severe renal insufficiency in an attempt to delay the use of either CYA or TAC, which may be associated with worsening of renal function (41).

IL-2 Receptor Blockers

T-cell activation and proliferation following presentation of a foreign antigen requires the induction of several cytokines, including IL-2 (interleukin 2). Antibodies directed against the interleukin (IL)-2 receptor are effective for initial immunosuppression, as IL-2 receptor blockade down regulates IL-2 mediated T-cell proliferation. The IL-2 receptor antibodies such as Basiliximab and Daclizumab, given intravenously at the time of transplant and during the first posttransplantation week can reduce the incidence of acute liver graft rejection when utilized in combination with a calcineurin inhibitor, although these agents may not be sufficient to prevent rejection when utilized alone. The IL-2 receptor antibodies are generally well tolerated, although side effects may include infections, gastrointestinal distress, and rarely, pulmonary edema and bronchospasm. As these agents rarely induce renal dysfunction, many transplant programs utilize IL-2 receptor antibodies as "induction" therapy in individuals with renal insufficiency at the time of transplantation (42), in an attempt to delay initiation or diminish dose of calcineurin inhibitors, which may exacerbate renal insufficiency.

Calcineurin Inhibitors

IL-2 inhibition effectively suppresses T-Cell activation. Cyclosporine and TAC achieve this by binding to cytoplasmic receptors, forming complexes which inactivate calcineurin, a key enzyme in T-cell signaling. The major side effects of both CYA and TAC include hypertension, renal insufficiency, and neurologic complications. However, there is evidence to suggest that obesity, hyperlipidemia, hirsutism, and gingival hyperplasia occur more commonly in patients who receive CYA, while a higher rate of diarrhea, insulin resistance, and diabetes is seen in patients who receive TAC. In response to inconsistent absorption of standard Cyclosporine, the development of a microemulsified formulation of cyclosporine (e.g., Neoral) has allowed consistent blood levels (43). Given their efficacy and oral administration, calcineurin inhibitors have a central role in posttransplant immunosuppression.

Safety and efficacy of calcineurin inhibitors is generally assessed by monitoring blood levels drawn prior to the dose (trough), although several investigators describe that blood levels drawn 2 h after a dose of Cyclosporine (i.e., C2 levels) rather than trough levels more accurately indicate exposure to drug. At many transplantation centers, the definition of appropriate target level of calcineurin inhibitor is linked to the patients time posttransplantation; in general, higher levels are required in the first several months postoperatively while the threat of rejection is acute. The target levels for calcineurin inhibitors can be appropriate adjusted downward as patients achieve both normal liver function and freedom from rejection months to years following surgery. In addition, a philosophy of minimizing exposure to high levels of calcineurin inhibitors in HBV or HCV infected patients is adopted by many transplant centers, due to the negative impact of "overimmunosuppression" on viral replication and disease recurrence.

Antiproliferative Agents

Antiproliferative agents such as AZA and MMF prevent the expansion of activated T cells and B cells and regulate immune mediated injury. Azathioprine, a purine analogue, is metabolized in the liver to its active compound, 6-mercaptopurine, which inhibits adenosine and guanine production, thus inhibiting DNA and RNA synthesis in rapidly proliferating T cells. Mycophenolate Mofetil is a potent noncompetitive inhibitor of inosine monophosphate dehydrogenase (IMPDH), an enzyme necessary for the synthesis of guanine, a purine nucleotide. Mycophenolate Mofetil, when used in combination with a calcineurin inhibitor and steroids has been shown to be associated with lower rejection rates in the first 6 months posttransplantation when compared to AZA (44). The major toxicities associated with the use of either MMF or AZA are bone marrow suppression with resultant leukopenia, anemia, and thrombocytopenia, though this is more marked with AZA. Mycophenolate Mofetil has been associated with a greater incidence of dyspepsia, peptic ulcers, and diarrhea when compared to AZA, while pancreatitis may occur in individuals prescribed AZA. These side effects usually abate by dose reduction or discontinuation. The majority of transplant centers utilize a combination of a Calcineurin inhibitor with either MMF or, less commonly, AZA for at least the first 3–6 months posttransplantation. Since AZA and MMF do not cause renal insufficiency, they can be utilized in a strategy to minimize or avoid calcineurin inhibitor use, particularly in patients with renal dysfunction.

Other Immunosuppressive Agents

The limitations and untoward effects of available immunosuppressive agents have induced research and development of alternative agents. Sirolimus (Rapamycin) (RAPA) and its derivative Everolimus represent a new class of compounds, which achieve their immuosuppressive effect by inhibiting mTOR (target of Rapamycin). Inhibition of mTOR diminishes intracellular signaling distal to the IL-2 receptor and prevents T-cell replication. As the lymphoproliferative pathways inhibited by RAPA and Everolimus are distinct from those affected by calcineurin inhibitors, investigators have utilized these agents in combination with calcineurin inhibitors to achieve synergistic effect. However, enthusiasm for RAPA has been tempered by recent data showing higher rates of hepatic arterial thrombosis in patients who receive RAPA in the weeks immediately following transplantation (45). In addition, impaired wound healing has been noted in patients who receive RAPA, potentially due to impairment of granulation mediated by inhibition of TGF-β. Leukopenia, thrombocytopenia, and hyperlipidemia are the principal toxicities associated with RAPA and Everolimus. Recent reports of pneumonitis in RAPA treated patients have also emerged. A positive attribute of both RAPA and Everolimus is the absence of renal toxicity; some data suggest that post transplantation renal insufficiency can be reversed when calcineurin inhibitors are withdrawn and RAPA is initiated (46). Newer immunosuppressive agents will continue to be developed; it is hoped that these agents will be associated with diminished short- and long-term toxicity and facilitate a state of "immune tolerance" of the graft that will ultimately allow minimization of the requirement for immunosuppressive medications.

SUMMARY

Liver transplantation is the treatment of choice for appropriately selected patients with end stage liver disease.

Over the last several decades, significant advances in surgical technique and immunosuppression, selection of appropriate donors, grafts, and recipients, and improved therapies to prevent and treat postoperative complications have greatly improved posttransplantation outcomes. Despite these impressive achievements, many challenges remain. It is becoming increasingly apparent that the growing disparity between the number of liver transplant candidates and available organs will be associated with escalating death rates on the transplant waiting list. Enhanced posttransplantation survival has led to the emergence of complications associated with patient longevity, including nonhepatic disease, complications of immunosuppression, infections, neoplasia, and recurrence of the primary disease for which the liver transplantation was indicated. Further progress in liver transplantation will be achieved by maximizing the use of available organs, refinement and exploration of alternatives to deceased donor liver transplantation, improvements in immunosuppression, and enhanced recognition and treatment of long-term complications, particularly recurrent liver disease.

BIBLIOGRAPHY

1. Welch CS. A note on transplantation of the whole liver in dogs. Transplant Bull 1955;2:54.
2. Kukral JC, Littlejohn MH, Williams RK, Pancer RJ, Butz GW Jr, Starzl TE. Hepatic function after canine liver transplantation. Arch Surg 1962;85:157–165.
3. Starzl TE, Marchioro TL, Porter KA, Brettschneider L. Related Articles, Homotransplantation of the liver. Transplantation 1967;5(Suppl):790–803.
4. Mazzaferro V, Regalia E, Doci R, Andreola S, Pulvirenti A, Bozzetti F, Montalto F, Ammatuna M, Morabito A, Gennari L. Liver transplantation for the treatment of small hepatocellular carcinomas in patients with cirrhosis. N Engl J Med 1996;334: 693–699.
5. Moreno S, Fortun J, Quereda C, Moreno A, Perez-Elias MJ, Martin-Davila P, de Vicente E, Barcena R, Quijano Y, Garcia M, Nuno J. Martinez Liver transplantation in HIV-infected recipients. Liver Transpl 2004;11:76–81.
6. Kamath PS, Wiesner RH, Malinchoc M, Kremers W, Therneau TM, Kosberg CL, D'Amico G, Dickson ER, Kim WR. A model to predict survival in patients with endstage liver disease. Hepatology 2001;33:464–470.
7. Freeman RB, Wiesner RH, Edwards E, Harper A, Merion R, Wolfe R. United Network for Organ Sharing Organ Procurement and Transplantation Network Liver and Transplantation Committee. Results of the first year of the new liver allocation plan. Liver Transpl 2004;10:7–15.
8. Onaca NN, Levy MF, Sanchez EQ, Chinnakotla S, Fasola CG, Thomas MJ, Weinstein JS, Murray NG, Goldstein RM, Klintmalm GB. A correlation between the pretransplantation MELD score and mortality in the first two years after liver transplantation. Liver Transpl 2003;9:117–123.
9. Merion RM, Schaubel DE, Dykstra DM, Freeman RB, Port FK, Wolfe RA. The survival benefit of liver transplantation. Am J Transpl 2005;5:307–313.
10. http://www.UNOS.org.
11. Broelsch CE, Whitington PF, Emond JC, Heffron TG, Thistlethwaite JR, Stevens L, Piper J, Whitington SH, Lichtor JL. Liver transplantation in children from living related donors. Surgical techniques and results. Ann Surg 1991;214: 428–437.
12. Marcos A. Right-lobe living donor liver transplantation. Liver Transpl 2000;6:S59–S63.
13. Shiffman ML, Brown RS Jr., Olthoff KM, Everson G, Miller C, Siegler M, Hoofnagle JH. Living donor liver transplantation: summary of a conference at The National Institutes of Health. Liver Transpl 2002;8:174–188.
14. Emond JC, Renz JF, Ferrell LD, Rosenthal P, Lim RC, Roberts JP, Lake JR, Ascher NL. Functional analysis of grafts from living donors. Implications for the treatment of older recipients. Ann Surg 1996;224:544–552.
15. Otero A, Gomez-Gutierrez M, Suarez F, Arnal F, Fernandez-Garcia A, Aguirrezabalaga J, Garcia-Buitron J, Alvarez J, Manez R. Liver transplantation from maastricht category 2 non-heart-beating donors: A source to increase the donor pool? Transpl Proc 2004;36:747–750.
16. Ericzon BG, Larsson M, Herlenius G, Wilczek HE. Familial Amyloidotic Polyneuropathy World Transplant Registry. Report from the Familial Amyloidotic Polyneuropathy World Transplant Registry (FAPWTR) and the Domino Liver Transplant Registry (DLTR). Amyloid 2003;10:67–76.
17. Schemmer P, Mehrabi A, Kraus T, Sauer P, Gutt C, Uhl W, Buchler MW. New aspects on reperfusion injury to liver–impact of organ harvest. Nephrol Dial Transpl 2004;19:26–35.
18. Bhattacharjya S, Gunson BK, Mirza DF, Mayer DA, Buckels JA, McMaster P, Neuberger JM. Delayed hepatic artery thrombosis in adult orthotopic liver transplantation-a 12-year experience. Transplantation 2001;71:1592–1596.
19. Vignali C, Cioni R, Petruzzi P, Cicorelli A, Bargellini I, Perri M, Urbani L, Filipponi F, Bartolozzi C. Role of interventional radiology in the management of vascular complications after liver transplantation. Transpl Proc 2004;36:552–554.
20. Lefkowitch JH. Diagnostic issues in liver transplantation pathology. Clin Liver Dis 2002;6:555–570.
21. Fondevila C, Ghobrial RM, Fuster J, Bombuy E, Garcia-Valdecasas JC, Busuttil RW. Biliary complications after adult living donor liver transplantation. Transpl Proc 2003;35: 1902–1903.
22. Denys A, Chevallier P, Doenz F, Qanadli SD, Sommacale D, Gillet M, Schnyder P, Bessoud B. Interventional radiology in the management of complications after liver transplantation. Eur Radiol 2004;14:431–439.
23. Selzner N, Rudiger H, Graf R, Clavien PA. Protective strategies against ischemic injury of the liver. Gastroenterology 2003;125:917–936.
24. Clavien PA, Yadav S, Sindram D, Bentley RC. Protective effects of ischemic preconditioning for liver resection performed under inflow occlusion in humans. Ann Surg 2000;232: 155–162.
25. Demetris A, Adams D, Bellamy C, Blakolmer K, Clouston A, Dhillon AP, Fung J, Gouw A, Gustafsson B, Haga H, Harison D, Hart J, Hubscher S, Jaffe R, Khettry U, Lassman C, Lewin K, Martinez O, Nakazawa Y, Neil D, Pappo O, Parizhskaya M, Randhawa P, Rasoul-Rockenschaub S, Reinholt F, Reynes M, Robert M, Tsamandas A, Wanless I, Wiesner R, Wernerson A, Wrba F, Wyatt J, Yamabe H. Update of the international banff schema for liver allograft rejection: Working recommendations for the histopathologic staging and reporting of chronic rejection. An International Panel. Hepatology 2000 Mar; 31(3):792–799.
26. Neuberger J. Recurrent primary biliary cirrhosis. Baillieres Best Pract. Res Clin Gastroenterol 2000;14:669–680.
27. Wiesner RH. Liver transplantation for primary sclerosing cholangitis: timing, outcome, impact of inflammatory bowel disease and recurrence of disease. Best Pract Res Clin Gastroenterol 2001;15:667–680.
28. Molmenti EP, Netto GJ, Murray NG, Smith DM, Molmenti H, Crippin JS, Hoover TC, Jung G, Marubashi S, Sanchez EQ,

Gogel B, Levy MF, Goldstein RM, Fasola CG, Gonwa TA, Klintmalm GB. Incidence and recurrence of autoimmune/alloimmune hepatitis in liver transplant recipients. Liver Transpl 2002;8:519–526.

29. Burke A, Lucey MR. Non-alcoholic fatty liver disease, non-alcoholic steatohepatitis and orthotopic liver transplantation. Am J Transplant 2004;4:686–693.
30. Mackie J, Groves K, Hoyle A, Garcia C, Garcia R, Gunson B, Neuberger J. Orthotopic liver transplantation for alcoholic liver disease: A retrospective analysis of survival, recidivism, and risk factors predisposing to recidivism. Liver Transpl 2001; 7:418–427.
31. Berenguer M, Prieto M, Rayon J, Mora J, Pastor M, Vicente O, et al. Natural history of clinically compensated hepatitis C virus related graft cirrhosis after liver transplantation. Hepatology 2000;32:852–858.
32. Gane E. The natural history and outcome of liver transplantation in hepatitis C virusinfected recipients. Liver Transpl 2003;9:S28–S34.
33. Garcia-Retortillo M, Forns X, Llovet JM, Navasa M, Feliu A, Massaguer A, Bruguera M, Fuster J, Garcia-Valdecasas JC, Rimola A. Hepatitis C recurrence is more severe after living donor compared to cadaveric liver transplantation. Hepatology 2004;40:699–707.
34. Gaglio PJ, Malireddy S, Levitt BS, Lapointe-Rudow D, Lefkowitch J, Kinkhabwala M, Russo MW, Emond JC, Brown RS Jr. Increased risk of cholestatic hepatitis C in recipients of grafts from living versus cadaveric liver donors. Liver Transpl 2003;9:1028–1035.
35. Shiffman ML, Stravitz RT, Contos MJ, Mills AS, Sterling RK, Luketic VA, Sanyal AJ, Cotterell A, Maluf D, Posner MP, Fisher RA. Histologic recurrence of chronic hepatitis C virus in patients after living donor and deceased donor liver transplantation. Liver Transpl 2004;10:1248–1255.
36. Russo MW, Galanko J, Beavers K, Fried MW, Shrestha R. Patient and graft survival in hepatitis C recipients after adult living donor liver transplantation in the United States. Liver Transpl 2004;10:340–346.
37. Crippin JS, McCashland T, Terrault N, Sheiner P, Charlton MR. A pilot study of the tolerability and efficacy of antiviral therapy in hepatitis C virus-infected patients awaiting liver transplantation. Liver Transpl 2002;8:350–355.
38. Everson GT. Treatment of chronic hepatitis C in patients with decompensated cirrhosis. Rev Gastroenterol Disord 2004;4: S31–S38.
39. Gane E. Treatment of recurrent hepatitis C. Liver Transpl 2002;8:S28–S37.
40. Conti F, Morelon E, Calmus Y. Immuosuppressive therapy in liver transplantation. J Hepatol 2003;39:664–678.
41. Tector AJ, Fridell JA, Mangus RS, Shah A, Milgrom M, Kwo P, Chalasani N, Yoo H, Rouch D, Liangpunsakul S, Herring S, Lumeng L. Promising early results with immunosuppression using rabbit anti-thymocyte globulin and steroids with delayed introduction of tacrolimus in adult liver transplant recipients. Liver Transpl 2004;10:404–407.
42. Liu CL, Fan ST, Lo CM, Chan SC, Ng IO, Lai CL, Wong J. Interleukin 2 receptor antibody (basiliximab) for immunosuppressive induction therapy after liver transplantation: A protocol with early elimination of steroids and reduction of Tacrolimus dosage. Liver Transpl 2004;10:728–733.
43. Lilly LB, Grant D. Optimization of cyclosporine for liver transplantation. Transpl Proc 2004;36:267S–270S.
44. Wiesner R, Rabkin J, Klintmalm G, McDiarmid S, Langnas A, Punch J, McMaster P, Kalayoglu M, Levy G, Freeman R, Bismuth H, Neuhaus P, Mamelok R, Wang W. A randomized double-blind comparative study of mycophenolate mofetil and azathioprine in combination with cyclosporine and corticosteroids in primary liver transplant recipients. Liver Transpl 2001;7:442–450.
45. Trotter JF. Sirolimus in liver transplantation. Transplant Proc 2003;35:193–200.
46. Nair S, Eason J, Loss G. Sirolimus monotherapy in nephrotoxicity due to calcineurin inhibitors in liver transplant recipients. Liver Transpl 2003;9:126–129.

See also DIFFERENTIAL COUNTS, AUTOMATED; PHARMACOKINETICS AND PHARMACODYNAMICS; IMMUNOTHERAPY.

LONG BONE FRACTURE. See BONE UNUNITED FRACTURE AND SPINAL FUSION, ELECTRICAL TREATMENT OF.

LUNG MECHANICS. See RESPIRATORY MECHANICS AND GAS EXCHANGE.

LUNG PHYSIOLOGY. See PULMONARY PHYSIOLOGY.

LUNG SOUNDS

ROBERT G. LOUDON
RAYMOND L. H. MURPHY

INTRODUCTION

Medical devices and instrumentation have developed rapidly in the last few decades, yet the first diagnostic medical instrument, the stethoscope, is still the most widely used, and it has changed only superficially in design and function.

The lungs, as we breathe, produce sounds that are transmitted to the body surface and to the mouth. The characteristics of these sounds convey information about the sound-producing and -transmitting structures. This information often has diagnostic value. Auscultation of the lungs is therefore widely taught and practiced. Textbooks of physical diagnosis present a body of information that has been derived by careful workers since the introduction of the stethoscope by R.T.H. Laennec in 1819. Much of that information was indeed presented by Laennec himself in his remarkable treatise, *De l'Auscultation Mediate*(1,2).

In this article, the medical devices and instruments that have been applied to the study of lung sounds, including the traditional acoustic stethoscope are reviewed. This survey will include sound transducers and their placement, methods, and equipment used for the recording and analysis of lung sounds, results obtained by the use of these techniques, and their clinical meaning. Recent work on this subject helps in the understanding of what we hear with the stethoscope; some is aimed at answering specific questions in physiology or pathology, and some is designed to provide new diagnostic and monitoring tools. Much of this work has been done in the past three decades, reflecting the enormous increase in the availability and quality of sound recording and processing techniques during that period. Reviews of lung sounds (3–5) and the success of the International Lung Sounds Association and its annual meetings bear witness to the upsurge of interest in the subject. Recommended standards for terms and techniques used in computerized respiratory sound analysis (CORSA) have been prepared by a Task Force of the European Respiratory Society and published in the European Respiratory Review series (6). Better understanding of the meaning of current and future observations promises a larger place in the future for clinical and research applications.

THE STETHOSCOPE

The introduction of the stethoscope is an interesting story, well described in a bicentenary appreciation of Laennec's birth (7). Laennec, a young physician practicing in Paris, had occasionally found it useful to listen directly to a patient's chest, as had been done by physicians at least since the time of Hippocrates. In 1816, he wished to listen to an obese young lady's heart, but was reluctant to do so. He recollected (and this part of the story may be apocryphal) having seen boys playing on a park bench, one listening to the wooden bench at one end with his ear, and the other scratching the other end. Laennec's own words were that "he happened to recollect a simple and well know fact in acoustics, that sound could be transmitted through solid material or along a tube. He rolled a quire of paper into a sort of cylinder", placed one end over her heart, and listened at the other end. He was "not a little surprised and pleased" to hear the sounds more clearly in this "mediate" fashion than he had ever been able to do by the immediate application of his ear (2). Over the next 3 years he amassed an enormous amount of information about the sounds heard over the chests of his patients. As he did all of the autopsies at the Hopital Necker in Paris where he worked, he could often relate these sounds to the underlying pathology.

The first edition of Laennec's book (1) cost 13 francs for the two volumes; for an extra 2.50 francs, one received a wooden stethoscope. This "cylinder" served its purpose well. Modifications were introduced over the years, such as earpieces, flexible tubing, binaural stethoscopes, and a diaphragm on the chest piece. The relative merits of diaphragm and bell, the effect of the length and bore of the tubing, and the convenience of different patterns have been debated over the years, and the design of modern stethoscopes has been largely empirical, better models surviving because of their popularity with auscultators. Some characteristics that acousticians might think of as defects may indeed be advantageous from the physician's point of view. Those using them tend to feel comfortable listening to sounds with which they are familiar and may reject a stethoscope that lets them hear too much.

The assessment of acoustical performance of stethoscopes is not as simple as it might seem, and approaches to this problem have been described by several authors (8–10). The value of the traditional stethoscope is in no way reduced by the recent introduction of devices and instruments that can record and analyze the sounds that we hear. Rather, its value is increased. Appropriate use on new medical devices and instrumentation adds science to art, measurement to impression, and recordings to memory. Better understanding of what lung sounds mean, and of how much the simple stethoscope can tell us and how much it cannot, will make the use of the simple stethoscope in examining rooms or on clinical rounds more important than ever.

SOUND TRANSDUCERS

Microphones transform mechanical energy to electrical energy, in the sound frequency range. Mechanical movement at the chest wall, resulting from the transmission of vibrations representing lung sounds to the chest wall surface, may be detected by any one of several devices. The main categories are ceramic, condenser (capacitor), and electret microphones. Ceramic microphones use a piezoelectric ceramic element that produces voltage when it is stressed. They tend to be stable and rugged and do not need a bias voltage for operation. Condenser microphones of the conventional type act as a variable capacitor that requires a bias voltage. They have good sensitivity and frequency-response characteristics. Electret microphones are a more recent type; a permanent charge on the diaphragm and no free electrostatic charge on its surface relieve the need for a polarizing (bias) voltage and reduce sensitivity to humidity.

Most microphones are designed to receive sound transmitted through air. Air coupling has been used by several investigators recording sounds from the surface of the chest wall, or from the trachea, and it is not surprising that stethoscope chest-pieces have been used for this purpose. The sound transmitted through the air column in stethoscope tubing can be applied to a microphone just as it can to an auscultating eardrum. Direct mechanical coupling of the transducer to the signal site (chest wall or tracheal surface) is an alternative to air coupling.

Several authors have reviewed the relative advantages and disadvantages of the various types of microphones as lung sound transducers (11,12). Desirable characteristics include sensitivity, rejection of ambient noise and surface noise, appropriate frequency response, insensitivity to variation in pressure of application, ease of attachment, ruggedness, and low price. Sensitivity is necessary because of the low level of the sound signal. Vesicular breath sounds will on occasion be virtually inaudible, for example, when airflow rates at the mouth are <0.27 L/s (13).

It is not always possible to study lung sounds in ideal circumstances, and rejection of ambient noise is important for many applications. Microphone housing can be helpful in this regard. Heart sounds are often of greater amplitude than lung sounds and may obscure them. They can be made less troublesome by the frequency response of the microphone because heart sounds are in a lower frequency range. Air coupling or inherent microphone characteristics may help by increasing the high frequency response. Microphone placement can also reduce the interference from heart sounds, which are, of course, loudest over the front of the chest, particularly in the left lower zone, and are less obtrusive on the right side, especially at the base of the right lung posteriorly. One method that has been adopted to reduce contamination of lung sounds by the heart sounds is to record the electrocardiogram simultaneously and to use some form of gating to delete segments where the heart sounds are present (14). The periodicity of the heart sound makes this an attractive alternative for some purposes. Muscle noise can also contaminate lung sounds; again the frequency content of muscle noise is considerably lower than that of lung sounds, and a microphone that is insensitive to low frequency noise, or subsequent filtration of the signal, can be helpful. Muscle noise has the disadvantage of being timed with respiration because it arises from respiratory muscle activity, and this prevents it from being gated out on a time base. Most investigators have found that the frequency range of most interest in the recording and analysis of lung sounds lies between 100 and 1000 Hz, well within the frequency range of most microphones.

Surface noise is another important source of difficulty that can arise in recording and interpreting lung sounds. The movements associated with respiration make it easy for the microphone to slide over the skin surface in phase with respiration, producing sounds that are in phase with respiration, may be in the same frequency range as lung sounds, and may be very difficult to distinguish from friction sounds such as a pleural friction rub. Surface noise is more likely to arise when the microphone is mechanically in contact with, but not firmly fixed to, the chest wall. Air coupling may have advantages over mechanical coupling in this respect, but not always if the chest piece is of the diaphragm type commonly used in stethoscopes. Respiratory movement may also cause changes in the pressure with which a microphone is applied to the chest wall; if the microphone is strapped to the chest by a circumferential band, pressure on the microphone will increase as inspiration occurs and the chest diameter increases. Variation in pressure of the microphone against the chest wall is liable to alter the acoustic coupling, particularly if mechanical coupling is used to transmit surface movement to the sensitive microphone element. If the pressure exerted is sufficient, the deformation of the sensitive element may approach the limit of its range, damping the signal. Air-coupled microphones are less sensitive to changes in pressure of application, provided that the air chamber between chest wall surface and microphone element is vented to theoutside, usually by a small-bore needle; but too large a vent may increase the amount of ambient sound recorded (15).

For some purposes, the sound transducer is applied only briefly at a specific site on the chest wall while a few breaths are recorded. For monitoring purposes, attachment of a sound recording device for a period of hours or overnight may be necessary. Lightness and small bulk are important in this type of application, and in some cases two-sided adhesive tape or an adhesive patch similar to that used for electrocardiograph electrodes is adequate for attachment.

If chest wall surface movement is unimpeded, the vibrations that correspond to the lung sound do not involve actual mechanical displacement of the chest wall surface by more than a few micrometers. A sensor applied to the surface may measure displacement or, if it applies a load to the chest wall surface, it may measure pressure rather than displacement, or a combination of the two. Some sound transducers measure acceleration rather than actual physical displacement. In each case, the reaction of the sensor to the signal being sensed will influence its characteristics. Inertia, rigidity, or counterpressure by the sensing element may cause distortion of the sound. Particularly in the case of accelerometers, the mass of the sensing element will determine its frequency response characteristics. It is not always clear what criteria are used in making a decision about microphone type. The human ear is remarkably good at separating out the different sounds that may be combined to form a mixed signal, and often the final judgment may be made by listening to replay of a recorded signal. The efficiency of a particular sound system depends on the purpose for which it is intended, but unless the signal is listened to with an educated ear it is easy to be misled by, for example, frequency components whose origin is not obvious from inspection of a graphic or calculated spectrum.

RECORDING AND DISPLAY SYSTEMS

Those using devices and instruments to study lung sounds will choose recording and display systems appropriate to their purpose. Audio tape and strip-chart recorders have now virtually all been replaced by computers or systems designed or modified for the purpose. The signals of interest may be presented to the observer audibly, visually, or in a variety of forms during and after analysis. Standard physical examination of the chest does, in a sense, present audible and visual displays to the clinician. The stethoscope presents an audible signal at the earpieces, and the clinician observes his patient breathe to get a visual display of respiratory movement.

For teaching purposes at the bedside, an electronic stethoscope or microphone may be connected to several headsets worn by students, by telemetry if preferred, giving the instructor an opportunity to share the sounds with them. In this way, a realistic learning experience is provided with less imposition on the patient's patience. Recording of sounds for teaching purposes usually involves a computer system, or an electronic stethoscope and audio tape recorder. Standard audiovisual equipment has been used for editing, for adding comments, and for preparation of cassettes or disks for distribution (16,17) for teaching purposes.

For research purposes, arrays of microphones are now available with computer recording, analysis, and display systems to show the distribution of sound signals over the surface of the chest (18–20). Brief differences in time of sound signals have clinical relevance by allowing comparison in timing of the same sound signal of a crackle or the start of a wheeze arriving at different surface sites in the same patient. And on a longer time base, in asthmatics, for example, the site, the frequency pattern, and the sound amplitude of wheezing may change during exercise, sleep, exposure to cold air or to inhalants such as pollen, or industrial exposure, or in response to drug treatment. Sleep disorders such as nocturnal asthma, the sleep apnea syndrome, and snoring, may be studied by sound monitoring. Nocturnal asthma and snoring are present in the same patient more often than would be expected as a result of chance alone, especially in asthmatics under the age of 40 (21). Snoring is a respiratory, but not a lung sound, as it rises in the upper airways, at or above the larynx. Possible explanations for the association with asthma, and sound monitoring methods and devices, have been reviewed (22).

Comparisons over long periods of time were once made by recording the results of analyses of wheezes, rather than by comparing the actual recorded sounds. The development and proliferation of computers with rapidly increasing audiovisual capability and storage capacity are now, however, changing the situation to allow storage of original data on tape or disk together with derived values. Kraman et al. (23) evaluated minidisk recorders, with their considerable increase of storage capacity for music, for lung sound recording. They found no distortion of frequency or waveforms that would interfere with this use. For some studies, analyzing sound signals in real time as they are being acquired makes it simpler to monitor results as they accrue, and helps direct the course of an experiment.

An early example of audiovisual recording is in a paper by Krumpe et al. (24), in which the authors discuss the evaluation of bronchial air leaks by auscultation and phonopneumography. They describe three patients who develop air leaks from the bronchi after resectional lung surgery and in whom "videophonopneumography" provided more precise correlation of abnormal sounds with the underlying visibly leaking bronchial abnormalities. Audiovisual tapes or disks are useful for teaching or demonstration purposes, by providing examples of classical or of unusual sounds.

Simultaneous sound recordings at several sites have been used to study the spatial distribution of lung sounds. This has provided information on regional ventilation, and on the localization of abnormalities in disease such as pneumonia, airways obstruction, bullae, or small areas of infarction, atelectasis, fibrosis, or interstitial lung disease. Indeed, lung imaging by sound production provides a potential alternative to chest X rays and computed tomography (CT) scans, without the need to inject possibly damaging energy or drugs.

For research purposes, analysis of sound signals and any associated physiological measurements were formerly conducted off-line. The signals were recorded on tape or disk and replayed for analysis. This allowed editing for selection of relevant segments of data and for quality control and signal conditioning, such as amplification, filtering, or attenuation. The purpose of each study will determine the equipment needs, but most current lung sound research uses computers with high speed audiovisual capabilities. These can be adapted to record lung sounds along with physiological respiratory variables, such as airflow, lung volume, and esophageal pressure, measured simultaneously, which can then be related to the lung sounds. If relationships in time are to be studied with any precision, it is necessary to record signals together on one medium and it is necessary to know the frequency characteristics of the items of equipment used, and the time delays introduced by filters, envelope detectors, integrators, frequency analyzers, and other acquisition or processing devices.

SOUND ANALYSIS

Sound amplitude and frequency content are the two measurements that most commonly form the basis of lung sound analysis systems. Early studies presented the sound signal as a time-amplitude plot. If such plots represent a respiratory cycle on a few centimeters of paper, the result is a compressed representation that superimposes many successive sound signal cycles to form an envelope. Simple integrating and rectifying circuits can provide the outline of the envelope as a single line, thus acting as an envelope detector, ac–dc converter, or sound-level meter. Filters incorporated in such circuitry can yield a method for comparing sound amplitude in different frequency bands (14,17) or to provide a signal believed to represent the important band range of vesicular sound from the ventilation point of view (25).

The sound spectrogram is really an extension of this principle, the signal of interest being passed repetitively through a narrow bandpass filter with slowly changing center frequency and the signals passed being assembled to present a graphic display of time on the horizontal axis, sound frequency on the vertical axis, and sound amplitude by the degree of blackening of the paper. Sound spectrograms of this type, used routinely in the speech sciences, were applied to heart and lung sounds extensively by McKusick et al. (26) and are still widely used to good effect.

The time-amplitude plot of a sound signal has been used to advantage in a different way by Murphy et al. (27). Features of the sound waveform cannot be studied in detail without using a rapid time sweep on an oscilloscope, and only a brief (a few milliseconds) segment can be viewed in this way. By digitizing a sound signal at a rapid rate and playing the signal back through a digital-to-analogue converter (DAC), a "time-expanded" waveform was prepared. This has proved of particular value in studying crackles (rales), the brief sounds heard over fibrotic, edematous, consolidated, or atelectatic lung. Measurable characteristics of these crackles, such as the initial or the largest deflection width, show diagnostic value and automatic methods for their measurement are now being applied.

The sound characteristics of rhonchi, as opposed to crackles (continuous versus discontinuous adventitious sounds) require an additional approach. Essentially, they are longer in duration, possessed of perceptible pitch, and have a repetitive waveform pattern. Waveform analysis is a rapidly moving field. Sound frequency spectrum analysis of lung sounds has most frequently been reported in terms of discrete Fourier analysis. Several workers have used a fast Fourier transform algorithm to measure frequency content of signal segments. One way of representing time-variant sound signals is to assemble a sequence of spectra with frequency on the horizontal axis, sound amplitude or power on the vertical axis, and time on an oblique axis. Usually, some overlapping of the sequential segments and appropriate windowing (e.g., Hanning) are used. The resulting "bird's-eye view" has proved to be readily related to sounds, providing a mental image that can evoke a mental image of the sounds represented. Individual peaks on a frequency spectrum may be related to individual wheezes coming from the chest, and peak detection programs have been used (28,29) to compare them statistically. The fast Fourier transform is the most frequently reported type of waveform analysis, but other techniques, such as those of linear predictive coding (LPC), the maximal entropy method of waveform analysis, fractal-dimension analysis, wavelet networks, and artificial neural networks, are being explored. They are most likely to prove useful in brief sounds, in timing the onset or rapid changes in complex sounds, or in noting time relationships among sounds recorded at separate or at adjacent sensors. Any graphic form of waveform analysis is more readily interpreted when it can be combined with visual examination of a simultaneous time-amplitude plot.

RESULTS AND CLINICAL APPLICATIONS

Increasing attention and techniques for more exact representation have led to a rapid growth in information available about lung sounds. The meaning of these various items of information will emerge more slowly, as will clinical applications. The objective, quantitative study of lung sounds, is still at an interesting rapid growth phase of development. It is clear that a good deal of information is contained in the signals that we hear emerging from the chest (2) and that auscultation is one of the safest of diagnostic procedures, since no external energy or chemical is inserted into the body. It is also clear that some of the information conveyed would be difficult to obtain in any other way. Much of it is regional or local and may be able to tell us about mechanical events and structural characteristics at specific sites in the chest (30,31). The vesicular lung sounds have been studied in sufficient detail that we now know more about the probable general range of bronchial dimensions involved in the production of these sounds, but not the exact site; the effects of flow rate and lung volume, but not the exact nature of the relationships; and we know that there are relationships between vesicular lung sound intensity and regional ventilation, but not their exact nature. The roles of production and of transmission of these sounds are not always easy to distinguish from one another in the end-product, sensed at the site of their detection, but recent work by Kiyokawa and Pasterkamp (32) shows progress in this distinction.

We know that wheezing indicates airflow obstruction and roughly its levels in the bronchial tree. We know that several factors, such as airway dimensions, geometry, and compressibility, are important. Endobronchial surface characteristics and the presence and nature of secretions may also have some effect. We know that flow rates and intrathoracic pressure and volume history affect wheezes; but we do not know the relative importance of these factors and the extent of variation from one disease state to another. Crackles are known to be associated with certain diseases and not with other radiographically similar diseases, but we are not sure why. We know that crackles from different types of abnormal lungs have different characteristics, but a great deal of clinical observation will be needed to test their diagnostic value: and physiological or pathological studies to understand the basic mechanisms involved.

Laennec's stethoscope—and for that matter the stethoscope pulled currently from the pocket of a white coat—allows the user to consider the sound of one breath at one place at one time. Medical devices and equipment are now being developed that can expand the observations in time, in space, in content, and in information; for example, from one or two breaths to hundreds of breaths, and from one specific point on the chest to the entire chest. From one breath described or remembered as vesicular, reduced in volume, with a few end-expiratory crackles the information may expand to an assembly of pages of tables and graphs showing a variety of measured features. These can include diagrams of the chest showing where and how the lungs and ventilation vary, where and how much airflow obstruction or lung collapse is present, and can offer a regional description of airways' diameters and other characteristics.

Que et al. (33) developed a system to measure tracheal flow from tracheal sounds, and to use this to estimate tidal volume, minute ventilation, respiratory frequency, mean inspiratory flow rate, and duty cycle. Careful observations and comparison of the results with simultaneously recorded pneumotachygraph-derived volumes in various postures allowed them to address the problems inherent in the adverse signal/noise ratio and the low level of the flow-derived sound at flow rates seen in quiet breathing. The system that they developed suggests that their method of phonospirometry measures overall ventilation reasonably accurately without mouthpiece, noseclip, or rigid postural constraints.

The study by Kiyokawa and Pasterkamp (32) in a sense complements this by measuring lung sounds at two closely spaced sensors on the chest surface. In five healthy subjects, volume-dependent variations in phase and amplitude of signals recorded over the lower lobe might reflect spatial variations of airways and diaphragm during breathing. These authors noted similar variations in phase and amplitude on passive sound transmission, suggesting that a difference in sound transmission was a more likely cause of the variations than a difference in sound generation. Their observations compare local sounds that reflect local circumstances; the observations discussed in the previous paragraph concern central sounds that reflect total ventilation.

Several systems are now available or under development that can record sound signals simultaneously from a number of sites, with or without associated physiological signals, and present the observations for read-out by the physician. Lung sound documentation and analysis can now be done on personal digital assistants (PDAs) as well as on laptop computers. Stethoscopes can be connected to these devices wirelessly or by a short cable. This allows objective quantification of these sounds at the bedside (34,35). A personal computer based "telemedicine" system has been described in which two remote hemodialysis sites were connected by high speed telephone lines to allow video and audio supervision of dialysis from a central site (36). Such equipment may eventually be used to supplement—or perhaps in some circumstances replace—other diagnostic devices such as fluoroscopy or other types of radiographic imaging. They have the great advantage of avoiding the subjection of a patient to any potentially harmful radiation or other energy, and can therefore be used over prolonged periods of time.

Transthoracic speed of sound introduced at the mouth or the supraclavicular space (35) can be mapped at several sites on the chest using sound input with specific characteristics. This may allow noninvasive monitoring of conditions such as pneumonia, congestive heart failure, or pleural effusion that increase intrathoracic density. Chronic obstructive lung disease may be detected by reading lung sound maps showing time intensity plots at several sites over the chest; this appears to be more accurate than current clinical diagnostic methods. The ability to detect diaphragmatic movement by multichannel lung sound analysis suggests that it may prove to be an inexpensive bedside test. It may also have useful applications in ventilator management.

It seems clear that wider application of these new developments in lung sound analysis will lead to safe, useful, and rewarding forms of clinical and physiological information that can answer many imaging, diagnostic, and monitoring problems.

BIBLIOGRAPHY

1. Laennec RTH. De l'auscultation mediate ou traite du diagnostic des maladies des poumon et du coeur, fonde principalement sur ce nouveau moyen d'exploration. 1st French ed., Volumes 2, Paris: Brosson et Chaude; 1819.
2. Laennec RTH, trans. by Forbes J 1st American ed. Philadelphia: James Webster; 1823. p 211.
3. Mikami R, Murao M, Cugell DW, Chretien J, Cole P, Meier-Sydow J, Murphy RL, Loudon RG. International symposium on lung sounds. Synopsis of proceedings. Chest 1987;92:342–345.
4. Bettencourt PE, Del Bono EA, Spiegelman D, Herzmark E, Murphy RL. Clinical utility of chest auscultation in common pulmonary diseases. Am J Respir Crit Care Med 1994;150: 1291–1297.
5. Pasterkamp H, Kraman SS, Wodicka GR. Respiratory sounds. Advances beyond the stethoscope. Am J Respir Crit Care Med 1997;156:974–987.
6. Sovijarvi AHA, Vanderschoot J, Earis JE. Computerized Respiratory Sound Analysis (CORSA): Recommended standards for terms and techniques. Eur Respir Rev 2000;10: 77:585–649.
7. Sakula A. Laennec RTH 1781–1926. His life and work. A bicentenary appreciation. Thorax 1981;36:81.
8. Ertel PY. Stethoscope acoustics and the engineer: Concepts and problems. J Audio Eng Soc 1971;19:182–188.
9. Charbonneau G, Sudraud M. Measurement of the frequency response of some commonly used stethoscopes. Consequences to cardiac and pulmonary auscultation. Bull Eur Physiol 1985;21:49–55.
10. Abella M, Formolo J, Penney DG. Comparison of the acoustic properties of six popular stethoscopes. J Acoust Soc Am 1992;91:2224–2228.
11. Charbonneau G, Racineux JL, Sudraud M, Tuchais E. An accurate recording system and its use in breath sounds spectral analysis. J Appl Physiol 1983;55:1120–1127.
12. Pasterkamp H, Kraman SS, DeFrain PD, Wodika GR. Measurement of respiratory acoustical signals. Comparison of sensors. Chest 1993;104:1518–1993.
13. Kraman SS. Lung sounds: Relative sites of origin and comparative amplitude in normal subjects. Lung 1983;161:57–64.
14. Pasterkamp H, Fenton R, Tal A, Chernick V. Interference of cardiovascular sounds with phonopneumography in children. Am Rev Respir Dis 1985;131:61–64.
15. Kraman SS, Wodicka GR, Oh Y, Pasterkamp H. Measurement of respiratory acoustic signals. Effect of microphone air cavity width, shape, and venting. Chest 1995;108:1004–1008.
16. Cugell DW. Use of tape recordings of respiratory sound and breathing pattern for instruction in pulmonary auscultation. Am Rev Respir Dis 1971;104:948–950.
17. Banaszak EF, Kory RC, Snider GL. Phonopneumography. Am Rev Respir Dis 1973;107:449–455.
18. Kompis M, Pasterkamp H, Wodicka GR. Acoustic imaging of the human chest. Chest 2001;120:1309–1321.
19. Sun X, Cheetam BM, Earis JE. Real time analysis of lung sounds. Technol Health Care 1998;6:11–22.
20. Bergstresser T, Ofengeim D, Vyshedskiy A, Shane J, Murphy R. Sound transmission in the lung as a function of lung volume. J Appl Physiol 2002;93:667–674.
21. Fitzpatrick MF, Martin K, Fossey E, Shapiro CM, Elton RA, Douglas NJ. Snoring, asthma and sleep disturbance in Britain: a community-based survey. Eur Respir J 1993;6:531–535.
22. Dalmasso F, Prota R. Snoring: analysis, measurement, clinical implications and applications. Eur Respir J 1996;9:146–159.
23. Kraman SS, Wodicka GR, Kiyokawa H, Pasterkamp H. Are minidisc recorders adequate for the study of respiratory sounds?. Biomed Instrum Technol 2002;36:177–182.
24. Krumpe PE, Hadley J, Marcum RA. Evaluation of bronchial air leaks by auscultation and phonopneumography. Chest 1984;85:777–781.
25. Ploysongsang Y, Martin RR, Ross RD, Loudon RG, Macklem PT. Breath sounds and regional ventilation. Am Rev Respir Dis 1977;116:187–199.
26. McKusick VA, Jenkins JT, Webb GN. The acoustic basis of the chest examination: Studies by means of sound spectrography. Am Rev Tuberc 1955;72:122–134.
27. Murphy RLH, Holford SK, Knowler WC. Visual lung sound characterization by time-expanded wave-form analysis. N Engl J Med 1977;296:968–971.
28. Baughman RP, Loudon RG. Lung sound analysis for continuous evaluation of airflow obstruction in asthma. Chest 1985;88:364–368.
29. Pasterkamp H, Tal H, Leahy F, Fenton R, Chernik V. The effect of anticholinergic treatment on post exertional wheezing in asthma studied by phonopneumography and spirometry. Am Rev Respir Dis 1985;132:16–21.
30. Nath AR, Capel LH. Inspiratory crackles and the mechanical events of breathing. Thorax 1974;29:695–698.
31. Murphy RLH, Jr., Gaensler EA, Holford SK, Delbono EA, Eppler G. Crackles in the early detection of asbestosis. Am Rev Respir Dis 1984;129:375–379.
32. Kiyokawa H, Pasterkamp H. Volume-dependent variations of regional lung sound, amplitude, and phase. J Appl Physiol 2002;93:1030–1038.
33. Que C-L, Kolmaga C, Durand L-G, Kelly SM, Macklem PT. Phonospirometry for noninvasive measurement of ventilation: Methodology and preliminary results. J Appl Physiol 2002;93:1515–1526.
34. Bergstresser T, Ofengeim D, Vyshedskiy A, Shane J, Murphy R. Sound transmission in the lung as a function of lung volume. J Appl Physiol 2002;93:667–674.
35. Paciej R, Vyshedskiy A, Shane J, Murphy R. Transpulmonary speed of sound input into the supraclavicular space. J Appl Physiol 2002;94:604–611.
36. Winchester JF, Tohme WG, Schulman KA, Collman J, Johnson A, Meissner MC, Rathore S, Khanafer N, Eisenberg JM, Mun SK. Hemodialysis patient management by telemedicine: Design and implementation. ASAIO J 1997;43:M763–766.

See also PULMONARY PHYSIOLOGY; RESPIRATORY MECHANICS AND GAS EXCHANGE.

LVDT. See LINEAR VARIABLE DIFFERENTIAL TRANSFORMERS.

MAB. See Monoclonal antibodies.

MAGNETIC RESONANCE IMAGING

W. F. Block
A. L. Alexander
S. B. Fain
M. E. Meyerend
C. J. Moran
S B Reeder
K. K. Vigen
O. Wieben
University of
Wisconsin–Madison
Milwaukee
Madison, Wisconsin

J. H. Brittain
General Electric Healthcare
Milwaukee, Wisconsin

INTRODUCTION

The principle of nuclear magnetic resonance (NMR) was discovered by Felix Bloch and Edward Purcell independently in 1946. The two were awarded the Nobel Prize in Physics for the discovery, which had numerous applications in studying molecular structure and diffusion. Atomic nuclei with an odd number of protons or an odd number of neutrons behave like spinning particles, which, in turn, create a small nuclear spin angular momentum. This angular momentum of an electrically charged particle such as the nucleus of a proton leads to a magnetic dipole moment. In the absence of an external magnetic field, the orientation of these magnetic moments is random due to thermal random motion. These magnetic moments are referred to as spins, because the fundamentals of the phenomena can be explained using classical physics where the moments act similarly to toy tops or gyroscopes. The NMR phenomenon exists in several atoms and is used today to study metabolism via imaging. However, hydrogen is the simplest and most imaged nucleus in MR examinations of biological tissues because of its prevalence and high signal compared with other nuclei.

NMR imaging was renamed Magnetic resonance imaging (MRI) to remove the word nuclear, which the general public associated with ionizing radiation. MRI can be explained as the interaction of spins with three magnetic fields: a large static field referred to as B_0, which organizes the orientation of the spins; a radio frequency (RF) magnetic field referred to as B_1, which perturbs the spins so that a signal can be created; and spatially varying magnetic fields referred to as gradients, which encode the spatial location of the spins. These subsystems are shown in Fig. 1.

When an external magnetic field is present, the distribution of the magnetic moments is no longer random. Current technology allows large, homogenous static magnetic fields to be created using superconducting magnets, whereas smaller fields are possible with permanent magnets. In most conventional systems, the static field is aligned along the longitudinal axis or the long axis of the body, as shown in the z axis in Fig. 1. Clinical MRI scanners have been built with static fields ranging from 0.1 to 7 T, but the vast majority of scanners are between 0.5 and 3.0 T.

THEORY

Creating Net Magnetization

Consider a static field oriented along the z axis with magnitude B_0, or represented as a vector $\mathbf{B} = B_0\mathbf{k}$. Hydrogen protons have a quantum operator whose z component is quantized to ±½. According to quantum mechanics, only two discrete sets of orientations exist for the magnetic dipole of each hydrogen nucleus. In the parallel energy state, the magnetic moment vector μ orients itself so that its projection on the z axis aligns with the direction of the main magnetic field B_0. In the antiparallel energy state, this projection aligns in the opposite direction of the main field. It can be shown that the two allowed angles between magnetic dipoles and the static field are $\theta = \pm 54^\circ$ (1), and thus a population of spins will be oriented as in Fig. 2b.

The ratio of spins in the parallel state n^- to the spins in antiparallel state n^+ is given by the Boltzmann equation

$$\frac{n^-}{n^+} = e^{\frac{-\Delta E}{kT}} = e^{-\frac{\gamma 2\pi}{h}B_0 kT} \tag{1}$$

where γ is a nuclei-specific constant referred to as the gyromagnetic ratio, k denotes the Boltzmann constant, and T is the absolute temperature. There are only slightly more spins in the parallel state than in the antiparallel state because this state is of lower energy; however, the prevalence of water in biological tissue can create an adequate signal with this differential. This distribution of spin orientations in a small volume element results in an average or net magnetization $\mathbf{M}$, which aligns along the longitudinal or z axis, as shown in Fig. 2b. The entire process is referred to as polarization. The contributions in the transverse (x–y) plane sum to zero. As the argument of the exponential in Equation 1 is small and the difference in energy levels varies proportionally with field strength, the length of the net magnetization vector varies linearly with field strength. A quantum mechanics description of the spin distribution can be found elsewhere (2).

Signal Generation

The behavior of the net magnetization vector in an external field can be described by the classical model according to

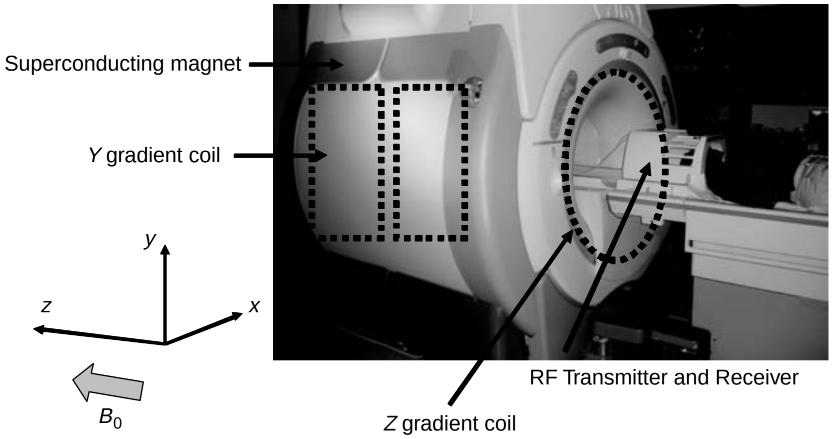


Figure 1. Clinical 1.5 T MRI scanner with static field oriented along long axis of the body (z). Patient's head lies in RF coil, which is used to both perturb and receive MR signal. Scanner bed will move patient into middle of cylinder before imaging begins. MR gradient coils for y dimension are shown, which have mirrored coils on the opposite side of the magnet. A portion of the z gradient, based on solenoid design, is also shown.

the Bloch equation.

$$\frac{d\mathbf{M}}{dt} = \gamma(\mathbf{M} \times \mathbf{B}) \tag{2}$$

A useful parallel description is a spinning toy top where the axis of the top is analogous to **M** and gravity is analogous to **B**. In the equilibrium state, the net magnetization **M** and the static magnetic field $\mathbf{B}_0$ are parallel so that **M** does not experience a torque and consequently the direction of **M** does not change. Similarly, the axis of a spinning top oriented vertically remains vertical.

The second magnetic field in MRI is an RF field that is created using an RF amplifier that supplies oscillating current into a coil that surrounds the patient. The coil is designed to create a magnetic field, referred to as $\mathbf{B}_1$ field, oriented in the transverse plane and approximately on the order of 10 T. By having the RFenergy oscillate at the resonant frequency of the nuclei, this relatively low field can perturb and rotate the net magnetization away from its orientation along the longitudinal axis. The resonant or Larmor frequency ω_0 is related to the static field strength such that $\omega_0 = \gamma B_0$. For protons, the gyromagnetic ratio $\gamma/2\pi = 42.57$ MHz/T. The field created by the tuned RF coil, referred to as an excitation, can be viewed as an applied torque that tips or flips spins away from the longitudinal axis by an angle referred to as the flip angle. The strength of the $\mathbf{B}_1$ field and the length of time it is applied determine the flip angle. The flip angle usually varies between 5 and 180° depending on the application.

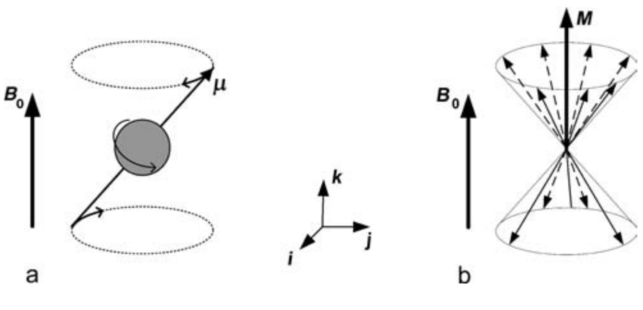


Figure 2. (a) shows the precession of a spin with a magnetic moment μ in a static field with magnetic flux density of B_0. An assembly of spins in parallel and antiparallel states is shown in (b). The Boltzmann equation determines the ratio of the spins in the two states. As the components in x and y compensate each other, the net magnetization **M** has a component in the z direction only (parallel to B_0). The coordinate system is shown with its unit vectors **i**, **j**, and **k** along x, y, and z.

Once the magnetization is no longer parallel to the static field, the right-hand side of Equation 5 is no longer zero and the direction of the net magnetization will change. In fact, it will begin to precess about the axis of the static magnetic field and at the Larmor frequency. In general, the precessional frequency is directly proportional to the magnetic field experienced by the spin, such that $\omega = \gamma B$. Similar to a toy top that is tipped an angle θ off its vertical axis, the top will maintain an angle of θ as it rotates about the vertical force supplied by gravity.

The net magnetization can be described by its longitudinal component M_z and its transverse component M_{xy}, a complex value whose magnitude describes the component's strength and whose angle describes the location of the component in the x–y plane. The rapid rotation of the transverse component will create a time-varying magnetic flux. A properly oriented receiver coil will detect this time-varying flux as a time-varying voltage, in agreement with Faraday's law of induction. Often, the same coil used for excitation can also be used for reception. This voltage signal, known as a Free Induction Decay, or FID, is shown after a 90° excitation in Fig. 3, which is the most basic form of a MR signal. Although the entire magnetization vector is tipped into the transverse plane in this example, smaller flip angles will also create a transverse magnetization and thus an FID.

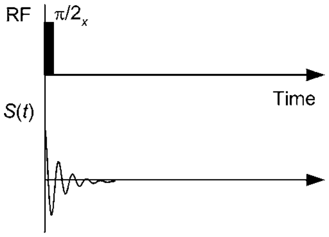


Figure 3. Generation of a free induction decay after a 90° RF excitation.

The complex motion of the net magnetization, and thus the recorded FID signal, can be described in a simplified manner by using a rotating reference frame that rotates at the Larmor frequency about the static B_0 field. In this rotating frame, the FID will decay as a simple exponential. The causes of this decay, and its use as potential image contrast mechanism, will be described after spatial encoding is described. Most MR signals are demodulating using the Larmor frequency and, thus are effectively acquired in the rotating frame.

Spatial Encoding

The first two magnetic fields described above allow biological tissue to be polarized, perturbed, and measured. In terms of clinical imaging, however, these fields merely allow us to integrate the signal derived throughout the body, a measure of little value. The field of MRI developed only when a third spatially varying magnetic field, referred to as a gradient field, was invented to spatially encode the MRI signal. This method allows us to achieve submillimeter resolution while using RF energy whose wavelengths are on the order of tens of centimeters to meters.

Dr. Paul Lauterbur realized, in 1973, that, instead of working like others to build a more homogenous field for NMR spectroscopy, spatially varying the strength of the magnetic field could provide a means to build an imaging system. For this work, he won the Nobel Prize in Medicine along with Sir Peter Mansfield in 2003.

The three gradient coils in a cylindrical MRI system, two of which are shown in Fig. 1, are laid out concentrically on a cylinder. The cylinder surrounds the patient as he or she lies inside of the static B_0 field. The three coils are designed to create longitudinal magnetic fields in z that vary in strength linearly with the x, y, and z dimensions, respectively. The digital scanner hardware controls the current waveforms, which are amplified by three respective gradient amplifiers before being sent to the gradient coils. The strength of each component gradient field, G_x,G_y, or G_z, is measured in G/cm or mT/m and is directly proportional to the current supplied to the coil. Changing gradient strengths quickly on clinical scanners is possible with amplifiers capable of slew rates of approximately 200 mT/m/s.

As the resonant frequency of an MR spin is directly proportional to the magnetic field it experiences, a gradient coil allows us to linearly vary the frequency of spins according to their position within the magnet. For example, a gradient of strength G_x, which does not vary in time, causes the frequency of spins to vary linearly with the x coordinate.

$$w(x) = \gamma[B_0 + G_x x] \tag{3}$$

Spins to the left of the magnet center rotate slower, spins at the exact magnet center remain unchanged, and spins to the right rotate faster than they did without the gradient.

Gradients can be used to selectively excite only spins from a slice or slab of tissue. Slice thicknesses in 2D MRI range from 1 to 20 mm. To select a transverse slice, the z gradient can be applied during RF excitation, which will cause the resonant frequency to vary as a function of z in the magnet, such that $w(z) = \gamma[B_0 + G_z z]$. Instead of exciting all the spins within the magnet, only spins whose frequency matches the narrow bandwidth of a pulsed-RF excitation will be excited within a slice at the center of the magnet. Modulating the frequency of the RF pulse up will move the slice superior in the body, whereas modulating it down will excite an inferior slice. Likewise, slices perpendicular to the x or y axis can be excited by applying a G_x or G_y gradient, respectively, simultaneously with RF excitation. In fact, an oblique slice orientation can be achieved with a combination of two or more gradients. The ability to control from which tissue signal is obtained, without any patient movement, is a major advantage of MRI.

Simplified MR Spatial Encoding. Once a slice of tissue is selected, the two remaining spatial dimensions must be encoded. A somewhat simplified method of visualizing encoding follows. For a transverse slice, receiver data could be obtained after RF excitation while a constant gradient was applied in the x direction. Tuning a receiver to select a very narrowband frequency range would determine which spins were present within a spatial range $x_1 < x < x_1 + \Delta x$. By repeating the experiment while changing the narrowband frequency range, a projection of the spin densities along the x axis could be determined. Likewise, the same experiment could be repeated while applying a constant y gradient to obtain a projection of spin densities along the y axis. Likewise, projections along arbitrary axes could be achieved by acquiring data while applying a combination of x and y gradients after RF excitation. In a matter very similar to computed tomography (CT) imaging, an image could be reconstructed from this set of acquired projections.

MR Spatial Encoding in the Fourier Domain. Although possible, the proposed method would be very slow because each sample point within each projection would require its own MR experiment or excitation. Time between excitations in MR vary in duration from 2 ms to 4 s depending on the desired image contrast. Even with the shortest excitation, each slice would require over 3 min of scan time. All the data for an entire projection can be acquired within milliseconds by considering how the phase of the transverse magnetization varies, instead of the frequency, with spatial position. This description also uses the concept that position in MR is encoding using an alternative Fourier domain where signal location is mapped onto spatial frequencies.

Integrating the frequency expression in Equation 3 indicates how the spin phase, or location of the transverse magnetization within the transverse plane, will vary with the x coordinate during a general time-varying gradient $G_x(t)$ applied after excitation. Ignoring the phase term due to the B_0 field, which will be removed during demodulation of the received signal, gives a phase term for each spin that varies with the spatial position x and the integral of the applied gradient at each point in time.

$$M_{xy}(x,t) = M_{xy}(x)e^{-j2\pi\frac{\gamma}{2\pi}\int_{t'=0}^{t}[Gx(t')x]dt'} \tag{4}$$

The signal received by the MR coil can be expressed as an integration of all the excited spins in the *x–y* plane using the following equation:

$$S(t) = \int_y \left[\int_x M_{xy}(x,y) e^{-j2\pi k_x(t)x} dx \right] dy \tag{5}$$

where a Fourier spatial frequency, termed $k_x(t)$ in MR, has been substituted for $\frac{\gamma}{2\pi}\int_{t'=o}^{t}(G_x(t')x)dt'$. In this example, the coil simply integrates all the spins in the *y* dimension, and thus only spatial information in *x* is available. The received signal can be seen as a 1D Fourier transform of the projection of transverse magnetization $M_{xy}(x, y)$ onto the *x* axis. The corresponding coordinate in the Fourier domain at each point in time *t* is determined by the ongoing integral of the gradient strength. Thus, we can acquire an entire projection in one experiment rather than numerous MR experiments as in the simplified example with the narrowband receiver.

The last spatial dimension for this 2D imaging example *y* has a corresponding Fourier dimension termed k_y, which can be similarly traversed by designing the integral of the G_y gradient current.

$$S(t) = \int_y \int_x M_{xy}(x,t) e^{-j2\pi k_x(t)x} e^{-j2\pi k_y(t)y} dxdy \tag{6}$$

where $k_y(t) = \frac{\gamma}{2\pi}\int_{t'=o}^{t}(G_x(t')x)dt'$. The integral of the gradients determines location in the Fourier space known as *k* space in MRI. Numerous strategies can be used to traverse and sample *k* space before transforming the data, often with a Fast Fourier Transform (FFT), into the image domain. The method can be extended to three dimensions by exciting a slab of tissue and using the G_z gradient to encode the third dimension.

As in the simplified example, a combination of G_x and G_y can be used to sample the 1D Fourier expression of projections of M_{xy} at arbitrary angles. This data can be transformed into the actual projections using 1D inverse Fourier transforms. Methods very similar to computed tomography can translate the projection data into an image. Although acquiring data in this manner, known as radial imaging, has interesting properties, by far the most popular method in clinical imaging traverses the Fourier space in a Cartesian raster pattern known as spin-warp imaging.

This sampling is typically completed in a series of experiments, where the time between consecutive excitations is referred to as the repetition time TR. In many MR acquisition schemes, a complete *k* space line is acquired along k_x, known as the frequency-encoding or readout direction, for each TR. During each subsequent TR, a line parallel to the previous one is sampled after applying a short, pulsed G_y gradient. By changing the strength of the pulsed G_y gradient by equal increments during each MR experiment, a different phase shift is placed on spins depending on their position in *y*. In terms of the *k* space formalism, the area under the G_y gradient pulse causes a vertical displacement in *k* space such that a different horizontal line in *k* space is acquired in each MR experiment, as shown in Fig. 4. Here, the vertical direction in *k* space is known as the phase-encoding direction. By applying 1D Fourier transforms in the k_x direction, spin position is resolved based on their frequency during readout. An image is formed by following these horizontal transforms with 1D Fourier transforms in the k_y dimension. Here, the *y* position of spins is resolved due to the different phase shifts experienced in each experiment prior to the readout gradient.

The image coverage, or field of view, in MRI decreases as the sampling rate decreases. As MR samples in the frequency domain, failure to sample fast enough in *k* space leads to aliasing in the image domain. Higher resolution in MRI requires obtaining higher spatial frequencies or larger extents of *k* space. Achieving adequate resolution and coverage then increases the amount of *k* space sampling that is required and increases imaging time. Unlike other modalities where hundreds to thousands of detectors can be used at a time, encoding spatial position in this method only allows one point of data to be taken at a time, which explains MR's relatively slow acquisition speed. Industrial scanners have only recently determined how to partially bypass this limitation by using up to 32 different receivers who have different spatial sensitivities to different tissues. The differences of each receiver in proximity, and thus sensitivity to each spin, can be used to synthesize unacquired regions of *k* space.

Image Contrast Through Varying Decay Rates

Imaging the spatial density distribution of hydrogen often produces a very low contrast image, as the density of hydrogen is relatively consistent in soft tissue. However, the imaging experiments described above can be easily modified to exploit the differences in time for which the

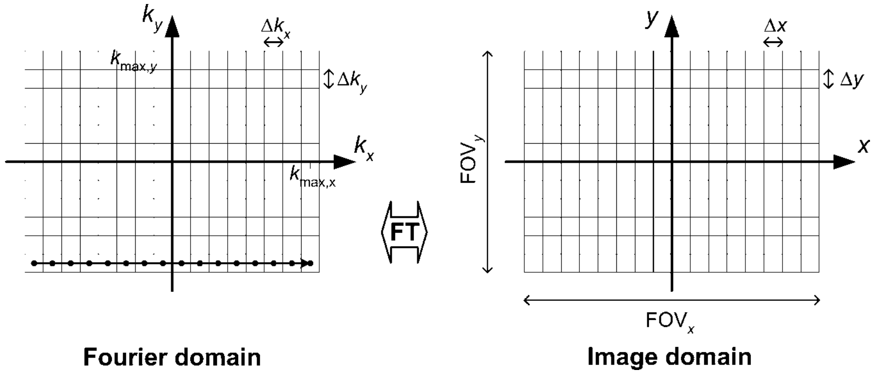


Figure 4. 2D spin-warp imaging with one readout per TR. One complete line is sampled along the readout direction on a rectilinear grid in the Fourier domain (*k* space) with a resolution of Δk_x (circles on the arrow). The next line is acquired parallel at a distance of Δk_y on the grid by increasing the phase-encoding gradient. This scheme is repeated until the desired grid is sampled. Images are reconstructed by an inverse 2D Fourier transform (FT).

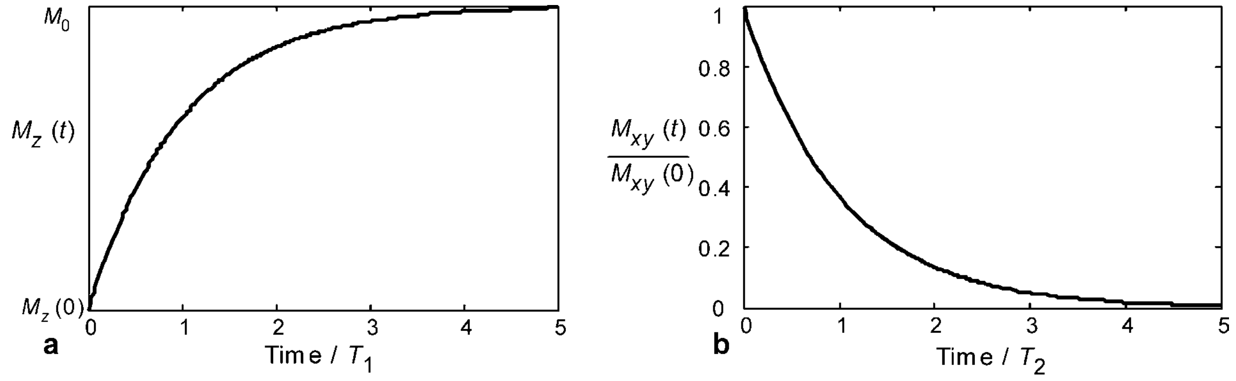


Figure 5. The regrowth of the longitudinal magnetization M_z (a) and the decay of the transverse magnetization M_{xy} (b) after an RF excitation.

MR spins remain perturbed. The differences account for the vast majority of image contrast in standard clinical MRI.

After the spins are perturbed, the transverse magnetization decays toward zero, whereas, the longitudinal magnetization returns toward its equilibrium magnetization. As more mechanisms exist for the loss of transverse magnetization than for the regrowth of longitudinal magnetization, the length of the magnetization vector **M** does not remain constant after excitation. Although related, the rate of longitudinal relaxation time, termed T_1, is always larger than the rate of transverse relaxation time, termed T_2.

If the magnetization has been completely tipped in the transverse plane with a flip angle of 90°, then the longitudinal magnetization recovers as

$$M_z = M_0[1 - e^{-t/T_1}] \tag{7}$$

T_1 is also called the spin–lattice relaxation time, because it depends on the properties of the nucleus and its interactions with its local environment. The transverse relaxation time T_2 is also referred to as the spin–spin relaxation time, reflecting dephasing due to interactions between neighboring nuclei.

$$M_{xy} = M_{xy}(0)e^{-t/T_2} \tag{8}$$

where $M_{xy}(0)$ is the initial transverse magnetization ($M_{xy}(0) = M_0$ for a 90° pulse). The temporal evolution of the longitudinal and transverse magnetization is shown in Fig. 5. In general, hydrogen protons in close proximity to macromolecules have lower relaxation times than bulk water that is freer to rotate and translate its position.

Delaying the encoding and acquisition of the transverse magnetization until some time after RF excitation generates T_2 image contrast. As injured and pathological tissues generally have higher T_2 relaxation rates, T_2-weighted images have positive image contrast. T_1-weighting can be achieved by using an interval between MR experiments, the TR parameter, which does not allow enough time for tissue to fully recover its longitudinal magnetization. Thus, tissues with shorter T_1 relaxation rates will recover more quickly and thus have more signal present in the subsequent experiments used to build the image than tissues with longer T_1. In general, T_1-weighting provides negative contrast for pathological tissue. An example is shown in Fig. 6 for a human brain tumor. The differing rates of

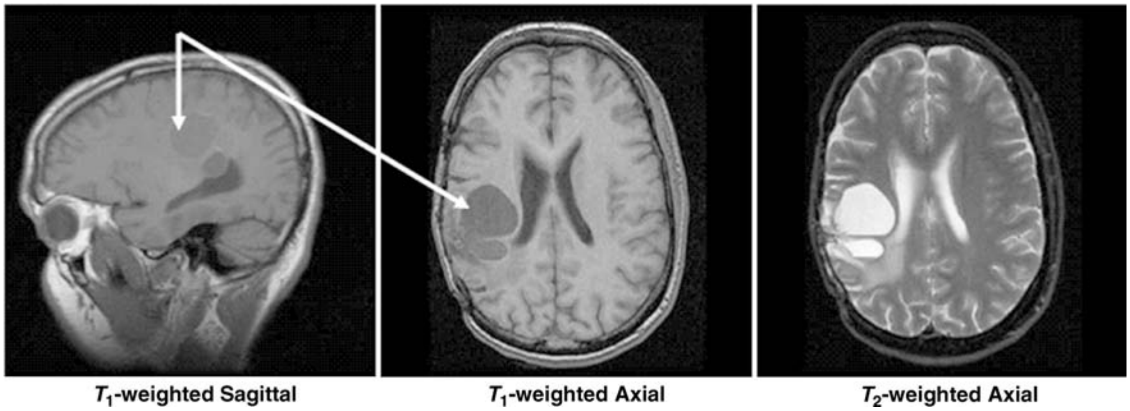


Figure 6. The flexibility of MR to image in different planes with different types of image contrast is shown in these sagittal and axial brain tumor (arrows) images.

Table 1. Longitudinal (T_1) and Transverse (T_2) Proton Nuclear Magnetic Resonance Relaxation Times for Several Tissues and Blood at 1.5 T

Tissue	T_1/ms	T_2/ms
Gray brain matter (3)	950	100
White brain matter (3)	600	80
Cerebrospinal fluid (CSF) (3)	4500	2200
Muscle (3)	900	50
Fatty tissue (3)	250	60
Oxygenated blood	1200	220
De-oxygenated blood	1200	120

recovery can also be used to null out an unwanted tissue, such as fat, by inverting all the spins 180° prior to imaging. As the point where unwanted tissue passes through the null of the recovery phase, an imaging experiment is begun. This technique is referred to as inversion recovery magnetization preparation or simply inversion recovery. Table 1 lists representative relaxation times for some tissues (3). Extensive reviews of the relaxations times (4) and methods for their measurement (3,4) are available.

Spin Echoes

Ideally, the transverse magnetization decays according to T_2. However, the signal dephasing in the transverse plane is significantly accelerated by field inhomogeneities due to difference in magnetic susceptibility between tissue types or the presence of paramagnetic iron. The largest inhomogeneities occur at air/tissue interfaces, such as near the sinuses or near the diaphragm. These effects lead to different precession frequencies and loss of coherence that are described by a T_2^* relaxation time

$$\frac{1}{T_2^*} = \frac{1}{T_2} + \frac{1}{T_2'} \tag{9}$$

where T'_2 represents the decay due to the effects described above.

A method to reverse these often undesirable dephasing effects uses a 90° pulse followed by a 180° spin refocusing pulse after a time delay τ_d, as shown in Fig. 7b. The first pulse rotates the longitudinal magnetization into the transverse plane as in the case of the FID. Prior to the second pulse, the magnetization dephases in the transverse plane due to T_2^* effects, with some spins rotating faster than the Larmor frequency and others spinning slower. The second pulse flips all magnetic moments about an axis in the transverse plane, effectively inverting the phase accruals due to different rotational frequencies. Over the second interval τ_d, the faster spins will catch up with the slower spins. As a result, a spin echo is said to form at the time $2\tau_d$, also known as the echo time or TE time. The amplitude of the signal at time TE is only decreased due to T_2 decay whereas the T_2^{*} effects have been reversed. A simplified pulse sequence for the generation of a spin echo is shown in Fig. 7b without the gradient waveforms necessary for spatial encoding.

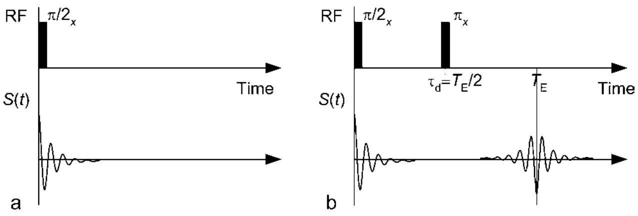


Figure 7. Generation of a (a) free induction decay and (b) a spin echo. In (a), other local factors dephase signal faster according to a T_2^* decay rate. If a second RF pulse is applied at time $\tau_d = \mathrm{TE}//2$ to flip the magnetization by 180°, the spins will refocus and form an echo at $TE = 2\tau_d$, which is only subject to T_2 decay.

Signal-to-Noise Ratios

Signal in MR is generally proportional to the number of nuclei and, thus, to the volume of the image voxel. Noise in MR is caused by the random fluctuations of electrons in the patient, and thus the source of noise is independent from the signal generating sources. The data acquisition system is designed such that the noise level from properly designed MR electronics will be dominated by patient noise. Overall, $SNR = \text{voxel volume} * \sqrt{\text{total data sampling time}}$.

Imaging Sequences

Ideally, all MRI would be performed with high spatial resolution, a high signal-to-noise ratio (SNR), ultrashort imaging time, and no artifacts. The difficulty in achieving all of these properties simultaneously has led to the development of many acquisition methods that differ in image contrast, acquisition speed, SNR, susceptibility to and type of artifacts, energy deposited in the imaged patient, and suppression of unwanted signal such as fat. Their corresponding images represent a combination of tissue-specific parameters T_1, T_2, proton density ρ, and scan-specific parameters such as repetition time (TR), echo time (TE), flip angle, field of view (FOV), spatial resolution, and magnetization preparation.

Gradient Recalled Echo (GRE) Imaging

Spin-echo imaging is desirable because signal voids due to magnetic field inhomogeneity are avoided that could mask pathological tissue or injury. Long repetition times, and thus long scan times, are necessary in spin-echo imaging to allow longitudinal magnetization to return after the relatively high flip angles used. Long scan times hinder the capture of dynamic processes such as the beating heart, cause discomfort to the patient, and limit the number of patients who can be imaged with this expensive resource. Thus, other methods of imaging have been developed. In gradient recalled echo (GRE) imaging, the echo is formed by dephasing and rephasing of the signal with gradient fields as shown in Fig. 8. In these diagrams, known as pulse sequence diagrams, plots of the time-varying gradient and RF waveforms are shown as function of time. Compared with the spin-echo sequences, gradient recalled echo imaging does not have a refocusing RF pulse. The absence of this pulse allows for a reduced minimal repetition time and echo time compared with spin-echo imaging, but the signal becomes susceptible to T_2^{*} decay rather than T_2 decay.

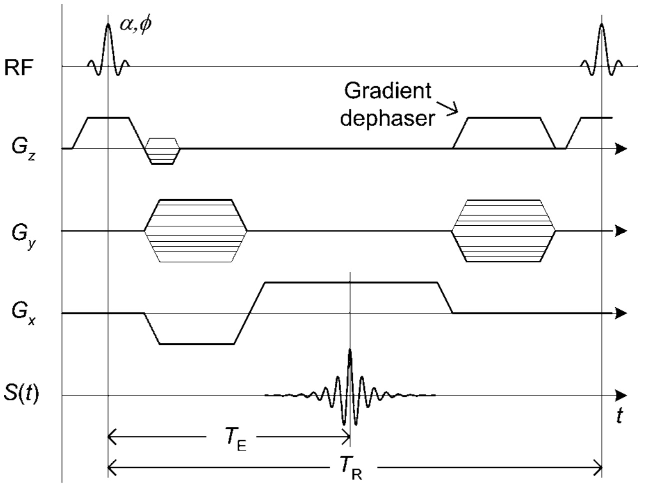


Figure 8. Basic 2D gradient echo pulse sequence. First, the magnetization is tipped into the transverse plane by an angle α during the application of a slice select gradient G_z. Then, the gradients G_y and G_x are used for phase encoding and the readout gradient. An echo forms at $t = T_E$ when the area under the readout gradient is zero. This experiment is then repeated every T_R with a different phase encode.

After RF excitation, the signal is dephased along the readout direction x with a prewinding gradient lobe. The amplitude of this gradient is then inverted to rephase the spins. When the area under the readout gradient is zero, the trajectory passes through the origin of k space and the echo forms with maximum amplitude. During the prewinder along the x axis, a gradient in y is played out to produce y depending on phase shifts for phase encoding.

By using a flip angle less than 90°, significant amounts of transverse magnetization are available without the need for long repetition times needed for recovering longitudinal magnetization. For example, after a single 30° excitation, the transverse magnetization contains sin(30°) or one-half of the available magnetization. Meanwhile, the longitudinal magnetization still contains cos(30°) or nearly 87% percent of the equilibrium magnetization. In fast GRE imaging, the repetition time is significantly reduced and generally less than the T_2 values of biological tissues. Under this condition, the transverse magnetization from preceding RF pulses is not completely dephased and generally results in a complex superposition of echoes from multiple RF pulses. Under certain conditions, a steady-state can be reached from repetition to repetition for one or more components of the magnetization (6).

GRE sequences can be used to generate T_1, T_1/T_2, T_2, T_2^*, and proton density-weighted contrast, depending on the choice of TR, TE, the flip angle α, and the phase ϕ of the RF pulse. By altering the phase of the RF transmit pulse in a pseudo-random method, the steady state of the transverse magnetization can be scrambled while the beneficial aspects of the longitudinal steady state are maintained. Although the signal from the transverse steady state is lost and only the signal from the current RF pulse is available, strongly T_1-weighted images are available with this technique, known as RF spoiling or spoiled gradient recalled (SPGR) imaging. This technique is popular with contrast-enhanced MR angiography, where an intravenously injected paramagnetic contrast agent significantly decreases the T_1 of blood while the T_1 of static tissues remains unchanged.

In an opposite approach, known as steady-state free precession (SSFP), the maximum amount of the transverse magnetization is maintained by rewinding all gradients prior to each RF pulse. The method provides T_2-like contrast very quickly and has proven very popular when fast imaging is essential such as in cardiac imaging.

Other Rapid MR Imaging Methods

In many applications, a short scan time is required to reduce artifacts from physiological motion or to observe dynamic processes. Many techniques exist to reduce the scan time while preserving high spatial resolution. One way to decrease spin-echo imaging time is to acquire multiple or all k space lines after a single preparation of the magnetization as explored with RARE (Rapid Acquisition with Relaxation Enhancement) (7). Also referred to as fast or turbo spin echo, the method works by creating a train of spin reversal echoes for which one line of k space is acquired for each. In echo-planar imaging (EPI) (8), an oscillating G_x gradient is used to quickly create many gradient echoes. By adding small blip gradients in between the negative and positive pulses of G_x, different horizontal lines in k space can be acquired.

Although the first MRI method proposed the acquisition of projections (9) as in CT, acquiring k space data on a Cartesian grid is fairly robust to magnetic field inhomogeneities and other system imperfections. Although spin-warp imaging (10) is predominant today, k space can be sampled along numerous 2D or 3D trajectories. The PROPELLER technique (11) acquires concentric rectangular strips that rotate around the origin, as shown in Fig. 8c. This method offers some valuable opportunities for motion correction due to the oversampling of the center of k space. K space can be sampled more efficiently with fewer echoes using spiral trajectories (12), as shown in Fig. 8d. Here, the amount of k space that can be acquired in one excitation is limited only by T_2 decay and possible blurring due to off-resonance spins. In nonCartesian acquisitions, phase errors due to off-resonance spins cause blurring. The sampling trajectories for these acquisitions schemes are shown in Fig. 9.

Applications

MRI is quickly moving beyond morphological and anatomical imaging. The advent of new functional image contrast mechanisms is making MR a tool for a much wider group of people than radiologists. Psychology, psychiatry, neurology, and cardiology are just some of the new areas where MR is being applied. A description of application areas in functional brain, diffusion-weighted brain, lung, MR angiography, cardiac, breast, and musculoskeletal imaging follows.

Functional Magnetic Resonance Imaging (fMRI).

Functional Magnetic Resonance Imaging (fMRI) is a method of

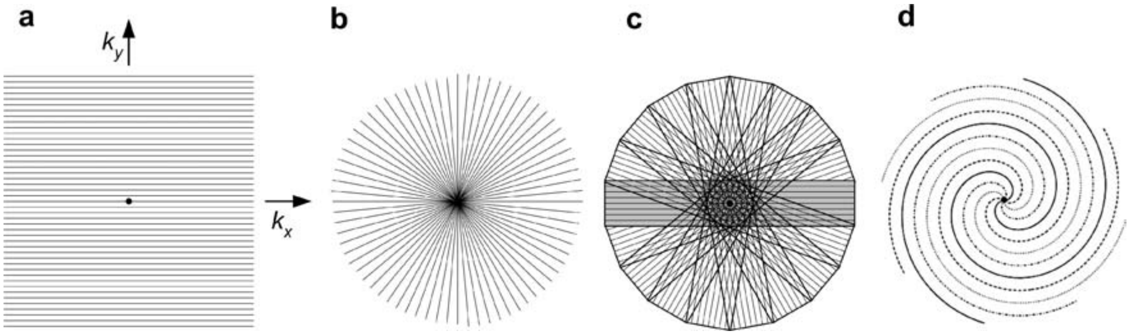


Figure 9. 2D k space sampling trajectories. Shown are the spin-warp (a), radial sampling (b), PROPELLER (c), and interleaved spiral imaging (d) examples.

measuring the flow of oxygenated blood in the brain (13–15). FMRI is based on the blood oxygen-level-dependent, or BOLD, effect. BOLD MRI is accomplished by first exposing a patient or volunteer to a stimulus or having them engage in a cognitive activity while acquiring single-shot images of their brain. The region of the brain that is responding to the stimulus or is engaged in the activity will experience an increase in metabolism. This metabolic increase will require additional oxygen. Therefore, an increase in oxygenated blood flow will occur (oxyhemoglobin) to the local brain area that is active. Oxyhemoglobin differs in its magnetic properties from deoxyhemoglobin. Oxyhemoglobin is diamagnetic like water and cellular tissue. Deoxyhemoglobin is more paramagnetic than tissue, so it produces a stronger MR interaction. These differences between oxyhemoglobin and deoxyhemoglobin in BOLD imaging are exploited by acquiring images during an "active" state (more oxyhemoglobin) and in a "resting" state (more deoxyhemoglobin), which creates a signal increase in the "active" state and a signal decrease in the resting state. Figure 10 shows a typical BOLD time course (shown in black) where four "active" states and four "resting" states exist. With prior knowledge of the activation timing (shown in red), we can perform a statistical test on the data to determine which areas of the brain are active. This statistical map (shown in color) is superimposed on a high resolution MR image so that one can visualize the functional information in relation to relevant anatomical landmarks.

Diffusion Imaging. The random motion of water molecules may cause the MRI signal intensity to decrease. The NMR signal attenuation from molecular diffusion was first observed more than a half century ago by Hahn (1950) (16).

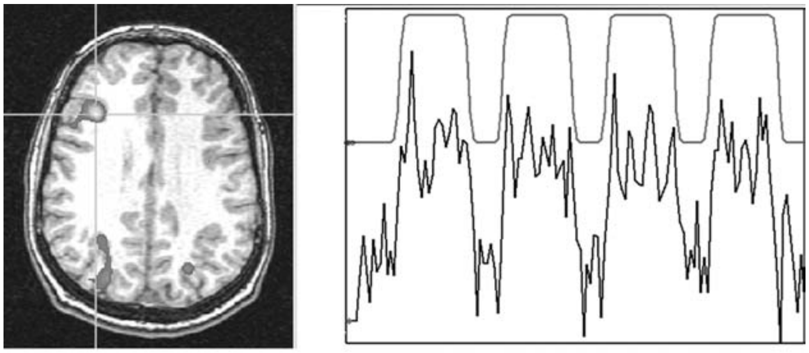
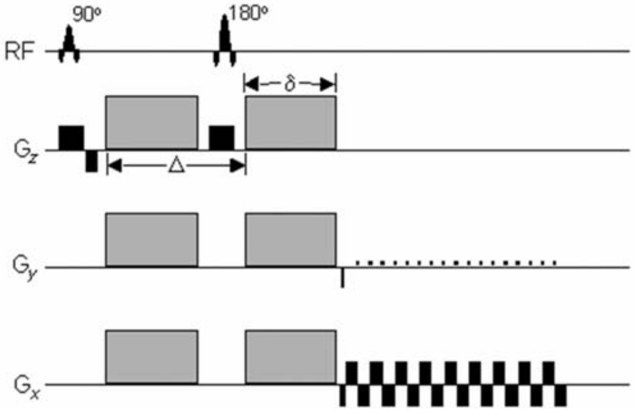


Figure 10. Color brain activation map is superimposed on high resolution MR image. Signal levels of the activated pixels are shown to increase during cognitive activity periods, whereas they fall off during periods of rest.



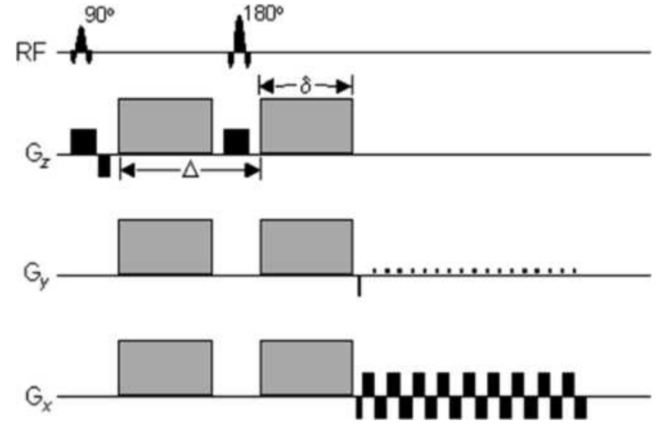


Figure 11. Temporal schematic of a diffusion-weighted, spin-echo pulse sequence with an EPI readout. The diffusion gradient pulses are shown as gray boxes on the gradient axes. The direction of diffusion-weighting can be changed by changing the relative weights of the diffusion gradients along G_x, G_y, and G_z.

Subsequently Stejskal and Tanner (1965) described the NMR signal attenuation in the presence of field gradients (17). More recently, field gradient pulses have been used to create diffusion-weighted MR images (18).

Diffusion-Weighted Pulse Sequences. Typically, diffusion-weighting is performed using two gradient pulses with equal magnitude and duration on each side of a 180° refocusing pulse, as shown in Fig. 11. The first gradient pulse dephases the magnetization as a function of position, and the second pulse rephases the magnetization. For stationary (e.g., no flow or diffusion) molecules, the phases induced by both gradient pulses will completely cancel, no signal attenuation will occur. In the case of motion in the direction of the applied gradient, a net phase difference will occur, $\Delta\phi = \gamma v G \delta \Delta$, which is proportional to the velocity v, the area of the gradient pulses defined by the amplitude G, and the duration δ, and the spacing between the pulses Δ. For the case of diffusion, the water molecules are also moving, but in arbitrary directions and with variable effective velocities. Thus, in the presence of diffusion gradients, the signal from each diffusing molecule will accumulate a different amount of phase, which, after summing over a voxel, will cause signal attenuation. For simple isotropic Gaussian diffusion, the signal attenuation for the diffusion gradient pulses in Fig. 11 is described by $S = S_o e^{-b\mathrm{D}}$ where S is the diffusion-weighted signal, S_o is the signal without any diffusion-weighting gradients (but otherwise identical imaging parameters), D is the apparent diffusion coefficient, and b is the diffusion-weighting described by the properties of the pulse pair $b = (\gamma G \delta)^2(\Delta\text{-}\delta/3)$.

Diffusion Tensor Imaging. The diffusion of water in fibrous tissues (e.g., white matter, nerves, and muscle) is anisotropic, which means the diffusion properties change as a function of direction. A convenient mathematical model of anisotropic diffusion is using the diffusion tensor (19), which uses a 3×3 matrix to describe diffusion using a general 3D multivariate normal distribution. The diffusion tensor matrix describes the magnitude, anisotropy, and orientation of the diffusion distribution. In a diffusion tensor imaging (DTI) experiment, six or more diffusion-weighted images are acquired along noncollinear diffusion gradient directions. Maps of the apparent diffusivity for each encoding direction are calculated by comparing the signal in an image without diffusion-weighting and the signal with diffusion-weighting. The diffusion tensor may then be estimated for each voxel, and maps of the mean diffusion, anisotropy, and orientation may be constructed, as shown in Fig. 12.

The primary clinical applications of diffusion-weighted imaging and DTI are ischemic stroke (20,21) and mapping the white matter anatomy relative to brain tumors and other lesions (22). DTI is also highly sensitive to subtle changes in tissue microstructure and, therefore, has become a popular tool for investigating changes or differences in the microstructure as a function of brain development and aging, as well as disease.

Vascular Imaging. Magnetic Resonance Angiography (MRA) describes a series of techniques that can be used to image vascular morphology and provide quantitative blood flow information in high detail. Two widely used techniques, phase contrast angiography and time-of-flight angiography, use the inherent properties of blood flow in the MR environment to create angiograms. A third technique, contrast-enhanced angiography, uses the injection of a

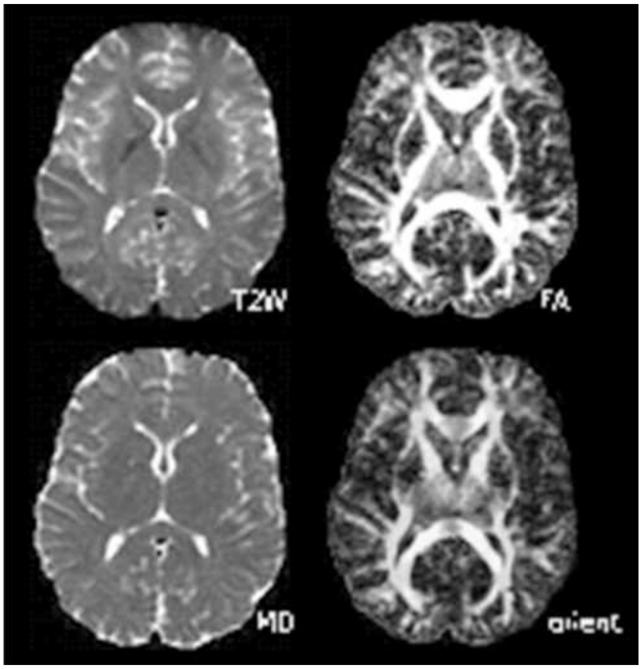


Figure 12. Representative diffusion tensor images. The images are (top-left): a T_2-weighted (or nondiffusion-weighted) image; (bottom-left): a mean diffusivity map (note similar contrast to T_2-weighted image with cerebral spinal fluid appearing hyperintense); (top-right): a fractional anisotropy map (hyperintense in white matter); and (bottom-right) the major eigenvector direction indicated by color (red = R/L, green = A/P, blue = S/I) weighted by the anisotropy (note that specific tract groups can be readily identified).

paramagnetic contrast agent into the vascular system to specifically alter the magnetic properties of the blood in relation to the surrounding tissue.

Phase-contrast (PC) angiography (23) usually uses a pair of gradient pulses of equal strength and opposite polarity, placed in the MRI sequence between the RF excitation pulse and the data acquisition window. During the imaging sequence, stationary nuclei accumulate phase during the first gradient pulse, and accumulate the opposite phase during the second gradient pulse, resulting in zero net phase. Moving nuclei accumulate phase during the first gradient pulse, but during the second pulse are in different positions, and accumulate phase different from that obtained during the first pulse. The net accumulated phase is proportional to the strength of the gradient pulses and the velocity of the nuclei. From the resulting data, images can be formed of both blood vessel morphology and blood flow.

TOF angiography techniques (24) (more accurately called "inflow" techniques) typically use a conventional gradient-echo sequence to acquire a thin 3D volume or a series of 2D slices. The nuclei in stationary tissue are excited by many consecutive slice-selective RF pulses. As a short TR is used, the longitudinal magnetization is not able to return to equilibrium, resulting in saturation of magnetization and low signal. Moving nuclei in the blood flow into the slice during each TR period, having been excited by zero or very few RF pulses. As these nuclei arrive in the imaging slice at or near full equilibrium magnetization, high signal is obtained from blood. Figure 13 shows a projection image of a 3D TOF dataset acquired in the head. The TOF technique can produce high quality MRA images in many situations, but slow or in-plane blood flow can result in blood signal saturation and reduced the quality of the images.

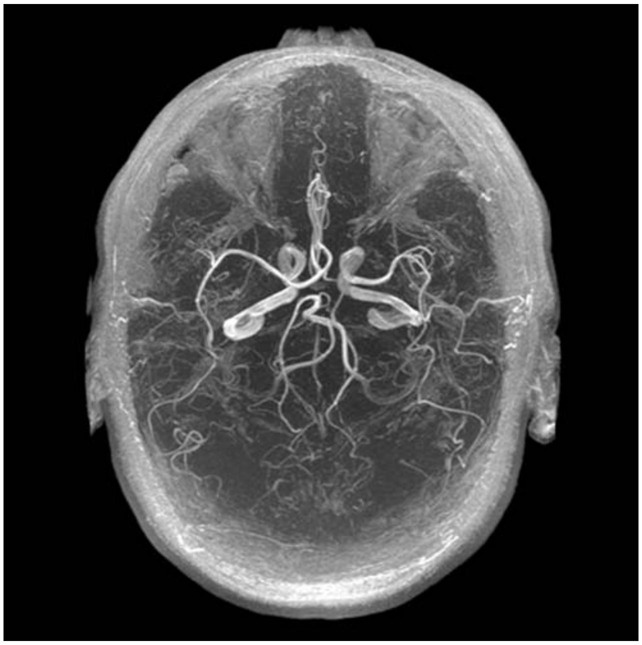


Figure 13. Time-of-flight (TOF) angiography in the head uses inflow of fresh blood to produce contrast between blood and the surrounding tissue. A Maximum Intensity Projection (MIP) reformatted image is used to compress the acquired volume data into a single slice for display.

Contrast-enhanced MRA (CE-MRA) is performed using an injection of a paramagnetic contrast agent into the intravenous bloodstream (25). Although several transition and rare-earth metal ions can be used, the most common is Gadolinium (Gd^{+3}) chelated to a biologically compatible molecule. The compound is paramagnetic, having a strong dipole moment and generating strong local magnetic field perturbations, which increase the transfer of energy between the excited hydrogen nuclei and the lattice, promoting T_1 relaxation and return to equilibrium of the longitudinal magnetization.

The contrast agent is injected intravenously in a limb away from the area of interest and circulates into the arterial system. The longitudinal relaxation rate is typically enhanced by a factor of 15 to 25 during this initial arterial phase, resulting in a much shorter T_1 for blood compared with the surrounding tissue. As the longitudinal magnetization in blood is much higher after each TR period, background tissue is suppressed in a manner similar to TOF imaging, and blood vessels have a comparably bright signal on the resulting images. Imaging is typically performed so that the central k space lines, which contain most of the image contrast information, are acquired while the contrast agent is distributed in the arteries of interest, but before it can circulate into the neighboring veins.

MRA data consist of large volumetric sets of image data, which are stored in the format of contiguous image slices. Specialized image display techniques are used to display the data in a manner that can be interpreted by the radiologist. Maximum Intensity Pixel (MIP) projections are widely used and are formed by projecting the volume set of data onto a single image plane. Here, each image pixel is obtained as the maximum value along the corresponding projection, as shown in Fig. 13. Volume rendering is beginning to be used more often to display MR angiograms. The individual slices of data are always available for detailed review by the radiologist and can be reformatted into any plane on the computer workstation to optimally display the vasculature of interest.

Cardiac MRI. Cardiac magnetic resonance (CMR) imaging is an evolving technique with the unprecedented ability to depict both detailed anatomy and detailed function of the myocardium with high spatial and temporal resolution. The past decade has seen tremendous development of phased array coil technology, ultra-fast imaging sequences, and parallel imaging techniques, all of which have facilitated ultra-fast imaging methods capable of capturing cardiac motion during breath-holding. The ability to perform imaging in arbitrary oblique planes, the lack of ionizing radiation, and the excellent soft tissue contrast of MR make it an ideal method for cardiac imaging. Comprehensive cardiac imaging is performed routinely in both in-patient and out-patient settings across the country and is widely considered the gold standard for clinical evaluation of many cardiac diseases (26).

Ischemic heart disease caused by atherosclerotic coronary artery disease (CAD) is the leading cause of mortality, morbidity, and disability in the United States, with over 7 million myocardial infarctions and 1 million deaths every year (27). Consequently, ischemic heart disease is the primary indication for CMR. Accurate visualization of wall thickness and global function (ejection fraction), as well as focal wall motion abnormalities, is performed with retrospectively ECG-gated ultra-fast short TR pulse sequences, especially steady-state gradient recalled echo imaging (28), as shown in Fig. 14. Breath-held cinemagraphic or CINE images have high SNR, excellent blood to myocardial contrast, and excellent temporal resolution (< 40–$50\,\text{ms}$) capable of detecting subtle wall motion abnormalities. Areas of myocardial infarction (nonviable tissue) are exquisitely depicted with inversion recovery (IR) RF-spoiled gradient echo imaging, acquired 10–20 minutes after intravenous injection of gadolinium contrast (29), as shown in Fig. 14. Areas of normal myocardium appear dark, whereas regions of nonviable myocardium appear bright (delayed hyper-enhancement). Accurate depiction of subtle myocardial infarction is possible because of good spatial resolution across the heart wall. The combination of motion and viability imaging is a powerful combination. Areas with wall motion abnormalities but without delayed hyper-enhancement may be injured or under-perfused from a critical coronary artery stenosis but are viable and may benefit from revascularization.

Cardiac "stress testing" using CMR has seen increasing use for the evaluation of hemodynamically significant coronary artery stenoses (30). Imaging of the heart during the first pass of a contrast bolus injection using rapid T_1-weighted RF-spoiled gradient echo sequences is a highly sensitive method for the detection of alterations

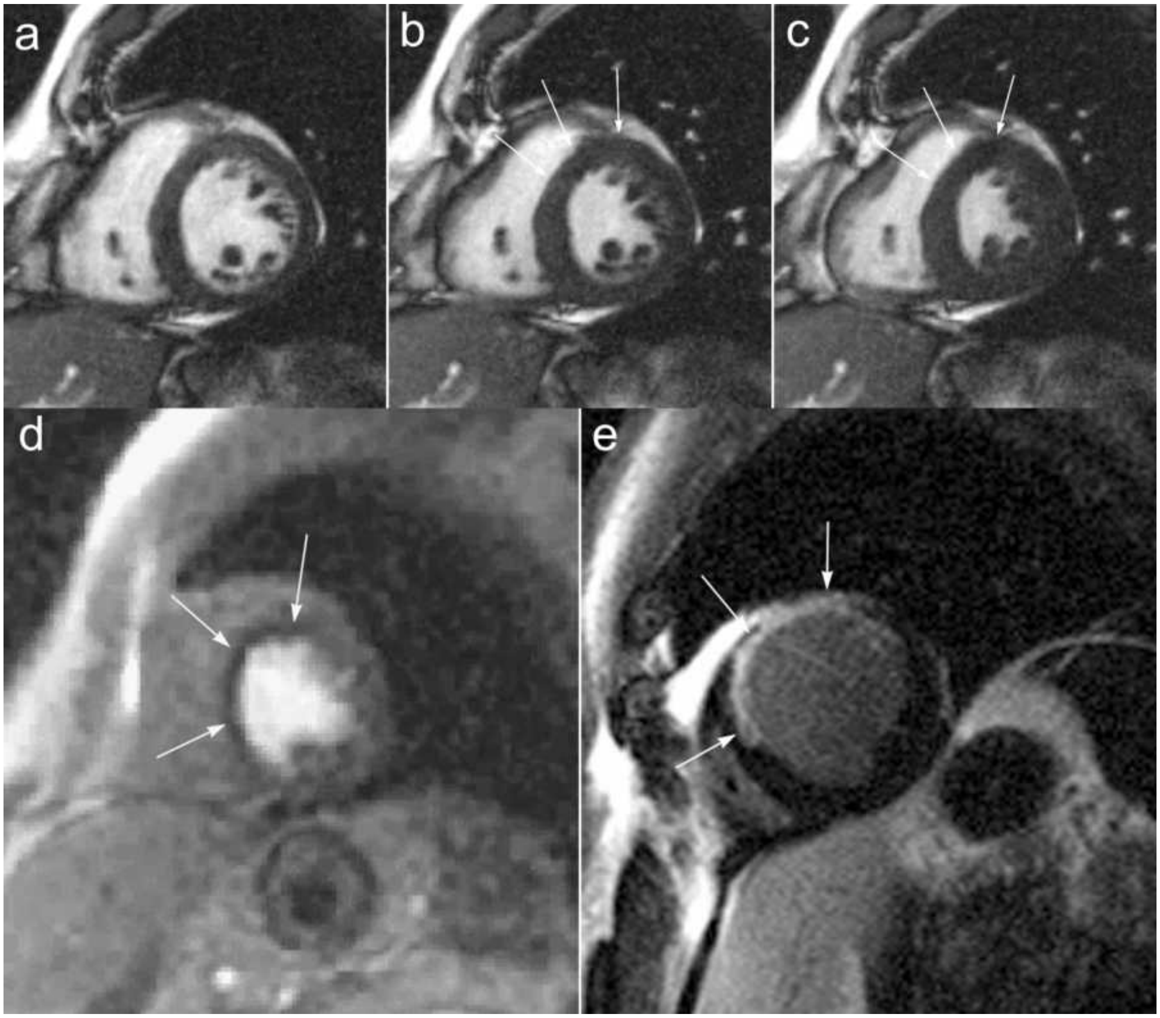


Figure 14. End-diastolic (a), mid-systolic (b), and end-systolic (c) short axis CINE images of the heart in a patient with a myocardial infarction in the anterior wall and septum, demonstrated by decreased wall thickening (arrows in b, c). The corresponding T_1-weighted RF-spoiled gradient recalled echo first-pass perfusion image (d) shows a fixed perfusion deficit (darker myocardium) in the corresponding territory (arrows). Finally, an inversion recovery RF-spoiled gradient echo image acquired at the same location (e) demonstrates a large region of delayed hyper-enhancement (arrows) indicating a full wall thickness region of nonviable myocardium that corresponds to the region of decreased perfusion and decreased contraction.

in myocardial blood flow (perfusion). Perfusion imaging during both stress (pharmacologically induced) and rest can reveal "reversible" perfusion defects that reflect a relative lack of perfusion during stress. In this way, coronary "reserve" can be evaluated and CAD can be uncovered, leading to further evaluation with coronary catheterization and possible angioplasty and stenting. Direct imaging of coronary arteries with CMR has shown tremendous technical advances, but is not commonly used, except for imaging of proximal coronary arteries in the evaluation of anomalous coronary arteries.

Other important indications of CMR include congenital heart disease, primarily, but not exclusively, in the pediatric population (31). Accurate diagnosis of a wide variety of congenital abnormalities requires high resolution, high contrast imaging that permits depiction of complex anatomical variants seen with congenital heart disease. Although anatomic imaging can be performed accurately with cardiac-gated CINE sequences, conventional sequences such as cardiac-gated black-blood fast spin-echo (FSE) and T_1-weighted spin-echo imaging are invaluable tools. Equally important to accurate anatomical imaging is functional imaging. With altered anatomy comes radically altered hemodynamics, requiring visualization of myocardial function with CINE imaging. Phase-contrast velocity imaging permits flow quantification through the heart, including the great vessels (pulmonary artery, aorta, etc.). An important example includes quantification of left-to-right "shunts" with resulting over-circulation of the pulmonary circulation. With a wide variety of pulse sequences, flexible scan plane prescription and the lack of ionizing radiation, CMR is ideally suited for evaluation of congenital heart disease.

Other important applications of CMR include visualization of valvular disease, pericardial disease, valvular disease, and cardiac masses. The latter two are particularly well evaluated with CMR; however, they are relatively uncommon and will not be discussed here.

Hyperpolarized Contrast Agents in MRI. Conventional MR imaging measures the resonant signal from the hydrogen nuclei of water, the most ubiquitous and highly concentrated component of the body. However, many other nuclei exist with magnetic dipole moments that produce MR signals. Many of these nuclei, such as phosphorous-31 and sodium-23, are biologically important in disease processes. However, these species typically exist at a very low concentration in the body, making them difficult to image with sufficient signal. One approach is to align, or polarize, the nuclei preferentially using physical processes other than the intrinsic magnetic field of the MR scanner. In some cases, these polarization processes can align many more nuclei than otherwise possible. These hyperpolarized nuclei can then act as contrast agents to better visualize blood vessels or lung airways on MRI. For example, helium-3 and xenon-129 are inert gases whose magnetic dipole moments can be hyperpolarized using spin-exchange optical pumping—a method of generating a preferred alignment of the nuclear dipoles using polarized laser light (32). As they are inert gases, polarized helium-3 and xenon-129 are used as inhaled contrast agents for visualizing the lung airspaces (upper-right panel) using MRI (33),(34). Unlike other parts of the body, conventional MRI of the lungs suffers from poor signal due to low water proton density and the multiple air-tissue interfaces that further degrade the MR signal in the upper left of Fig. 15. Hyperpolarized gas MRI has been particularly useful for depicting airway obstruction in several lung diseases including asthma (lower panel of Fig. 15) (35), emphysema (36), and cystic fibrosis (37). Additional techniques based on this technology show promise for MR imaging of blood vessels using injected xenon-129 dissolved in lipid emulsion (38), gas-filled microvesicles (39) and liquid-polarized carbon-13 (40). Hyperpolarized carbon-13 agents are of particular interest because of the wide range of biologically active carbon compounds in the body. Another important advantage of this technology is its ability to maintain high signal using low magnetic field (0.1–0.5 T) scanners (41). These systems are much cheaper to purchase and maintain than the high field (1.5–3.0 T) MRI scanners in common clinical use today.

Breast MRI. Breast MRI is presently used as an adjunct to mammography and ultrasound for the detection and diagnosis of breast cancer. Dynamic contrast-enhanced (DCE) MRI has been shown to have high sensitivity (83%–96%) to breast cancer but has also demonstrated variable levels of specificity (37–89%) (42). DCE-MRI requires an injection of a contrast agent and acquisition of a subsequent series of images to enable the analysis of the time course of contrast uptake in suspect lesions. Lesion morphology is also important in discerning benign from malignant lesions in breast MRI. Standard in-plane spatial resolution is sub-millimeter. A typical clinical breast MRI includes a spoiled gradient echo (SPGR) T_1-weighted sequence both precontrast (Fig. 16a) and, at minimum, at 30 second intervals postcontrast (Fig. 16b). Along with their morphologic characteristics, lesions can be further described by the shape of their contrast uptake curve. The three general categories of contrast uptake are (*1*) slow, constant contrast uptake (*2*) rapid uptake and subsequent plateau of contrast, and (*3*) rapid uptake and rapid washout of contrast (43). Although the slowly enhancing lesions are usually benign and fast uptake and washout is a strong indication of malignancy, time course lesion characterization is not absolute. The ambiguity of time course data for certain classes of lesions drives the investigation into higher temporal resolution imaging methods. A standard clinical breast MRI also includes acquisition of a T_2-weighted sequence for the identification of cysts (Fig. 16c). Present research in breast DCE-MRI is focused on development and application of pulse sequences that provide high temporal and spatial resolution. Also, investigation is ongoing into more specific characterization of uptake curves. Diffusion-weighted MRI, blood-oxygen-level-dependent imaging, and spectroscopy are also being investigated as possible methods to improve the specificity of DCE-MRI in the breast. In some circumstances, the high sensitivity of breast DCE-MRI outweighs the variable specificity leading to the present use of DCE-MRI to determine the extent of disease, with equivocal mammographic findings, and for the screening of high risk women.

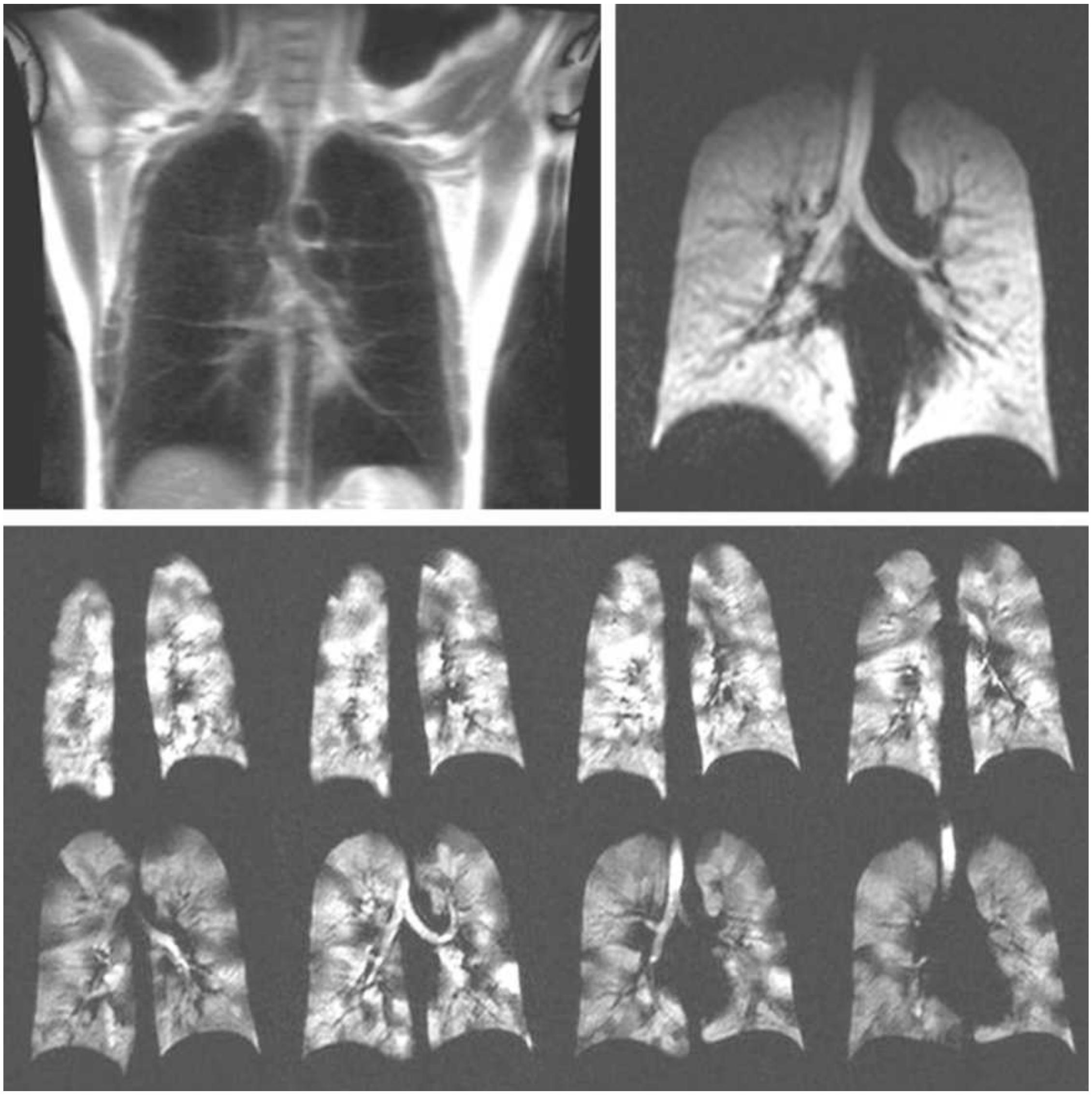


Figure 15. Upper left shows normal lack of signal in parenchyma of lungs in MRI. Ventilated areas are clearly seen after imaging inhaled hyperpolarized helium. Rapid imaging during inhalation and exhalation shows promise for capturing dynamic breathing processes.

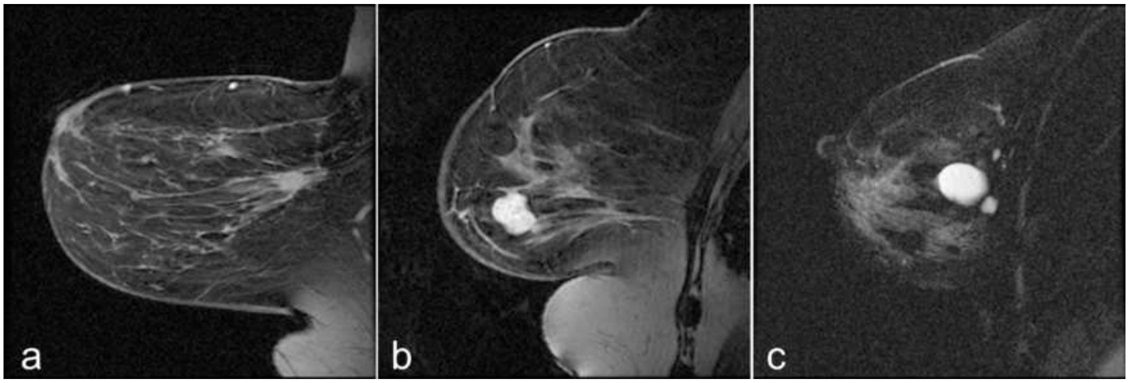


Figure 16. Fat-suppressed (a) pre-contrast T_1-weighted image (b) postcontrast T_1-weighted image (c), and noncontrast T_2- weighted image. Images are from different patients.

MRI of Musculoskeletal Disease. Musculoskeletal imaging studies traumatic injury, degenerative changes, tumors, and inflammatory conditions of the bones, tendons, ligaments, muscles, cartilage, and other structures of joints. Although X-ray imaging is the work-horse imaging modality for many musculoskeletal diseases, MRI plays a critical role in several aspects of diagnosis, staging, and treatment monitoring.

Fast spin-echo (FSE) pulse sequences are typically used to acquire T_1, T_2, and proton density-weighted images of the joints. For the assessment of joint structures, images are acquired in multiple planes to ensure adequate spatial resolution in all dimensions. These MR images can be used to evaluate tissues including ligaments, bone, cartilage, meniscus, and labrum. As a result of their high spatial resolution and excellent soft tissue contrast, MR images can provide accurate diagnosis that can prevent unnecessary surgeries and can facilitate pre-operative planning when surgical intervention is required.

Osteoarthritis is a degenerative disease that affects approximately 20 million Americans and countless others around the world. Currently, this debilitating disease is often not detected until the patient experiences pain that can be reflective of morphologic changes to joint cartilage. MR can be used to accurately measure cartilage thickness and volume. New MR techniques are also under development that may provide insight into biochemical changes in cartilage at the earlier stages of osteoarthritis that precede gross morphologic changes and patient pain.

Fortunately, primary bone tumors are relatively rare. However, bone is a common site for metastatic disease, which is especially true for breast, lung, prostate, kidney, and thyroid cancers. MR is a sensitive test for metastatic bone disease and is being adopted as a standard of care in some parts of the world, replacing nuclear scintigraphy. A typical approach employs inversion recovery pulse sequences to generate fat-suppressed, T_2-weighted images. Diffusion-weighted imaging also shows promise to detect hemotalogic cancers such as multiple myeloma, leukemia, and lymphoma.

Inflammatory diseases include infection and inflammatory forms of arthritis. Infection of the foot is a common complication of microvascular disease often seen with diabetes, a disease afflicting 18 million Americans. MR can be used to assess the vasculature of the foot as well as diagnose infection and evaluate treatment efficacy. Two million Americans have rheumatoid arthritis, a common form of inflammatory arthritis. Inflammation from a condition known as synovitis, which often occurs in rheumatoid arthritis patients, is shown in Fig. 17. New MR methods

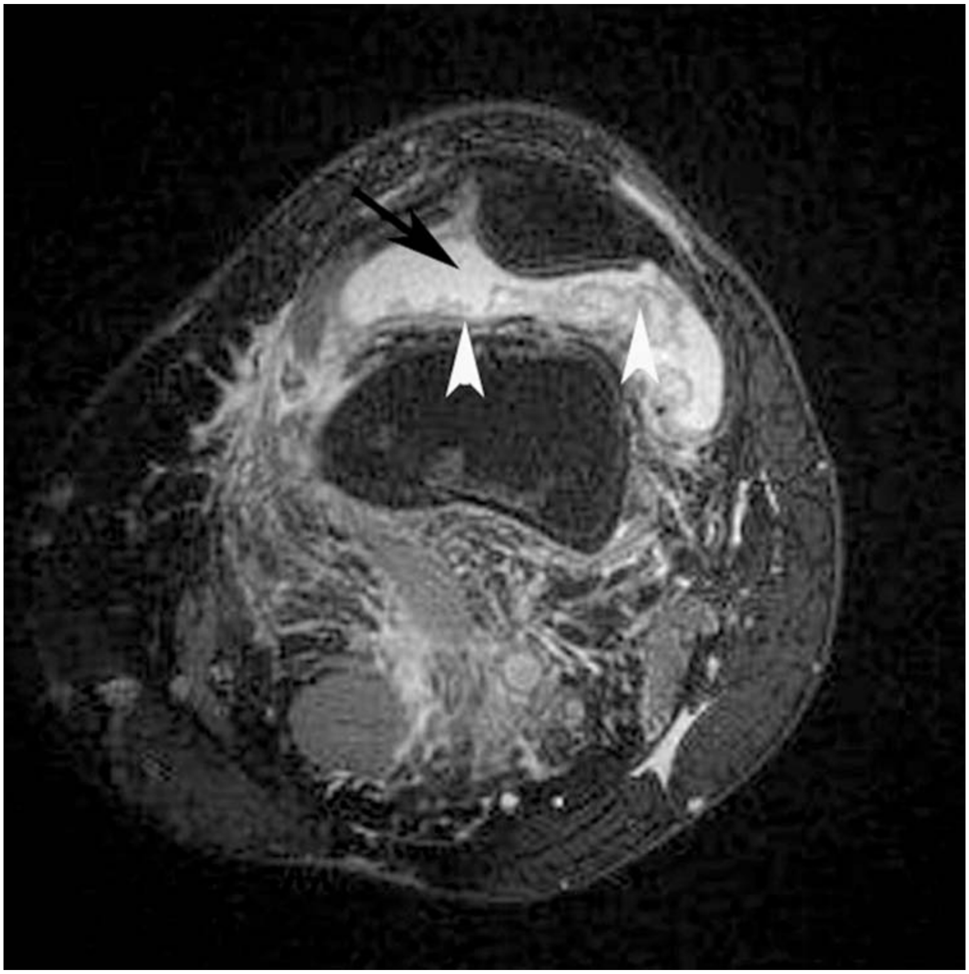


Figure 17. Axial knee image in a patient with inflamed synovium shows excellent soft tissue contrast available in MRI. The thickened intermediate signal intensity for synovium (arrowheads) is well distinguished from the adjacent high signal intensity of joint fluid (arrow).

are being developed to detect rheumatoid arthritis earlier and to gauge treatment success.

BIBLIOGRAPHY

1. Liang Z-P. Lauterbur PC. In: Akay M, ed. Principles of Magnetic Resonance Imaging—A Signal Processing Perspective. IEEE Press Series in Biomedical Engineering. New York: IEEE Press; 2000. p 416.
2. Abragam A. Principles of Nuclear Magnetism. Oxford, UK: Oxford University Press; 1994.
3. Haacke EM, et al. Magnetic Resonance Imaging-Physical Principles and Sequence Design. New York: Wiley; 1999. p 914.
4. Bottomley PA, et al. A review of H-1 nuclear-magnetic-resonance relaxation in pathology—are T1 and T2 diagnostic. Med Phys 1987;14(1):1–37.
5. Kingsley PB. Methods of measuring spin-lattice (T-1) relaxation times: An annotated bibliography. Concepts Magn Reson 1999;11(4):243–276.
6. Scheffler K. A pictorial description of steady-states in rapid magnetic resonance imaging. Concepts Magn Reson 1999; 11(5):291–304.
7. Hennig J, Nauerth A, Friedburg H. RARE imaging: A fast imaging method for clinical MR. Magn Reson Med 1986; 3(6):823–833.
8. Mansfield P. Multi-planar image formation using NMR spin echoes. J Phys C: Solid State Phys 1977;10:L55–L58.
9. Lauterbur PC. Image formations by induced local interactions: Examples employing nuclear magnetic resonance. Nature 1973;242:190–191.
10. Edelstein WA, et al. Spin warp NMR imaging and applications to human whole-body imaging. Phys Med Biol 1980;25(4):751–756.
11. Pipe JG. Motion correction with PROPELLER MRI: Application to head motion and free-breathing cardiac imaging. Magn Reson Med 1999;42(5):963–969.
12. Meyer CH. et al. Fast spiral coronary artery imaging. Magn Reson Med 1992;28(2):202–213.
13. Ogawa S, Lee T-M, Nayak A, Glynn P. Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. Magn Reson Med 1990;14:68–78.
14. Ogawa S, et al. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. Proc Natl Acad Sci USA 1990;87:9868–9872.
15. Bandettini PA, et al. Time course EPI of human brain function during task activation. Magn Reson Med 1992;25(2): 390–397.
16. Hahn E. Spin echoes. Phys Rev 1950;80(4):580–594.
17. Stejskal E, Tanner J. Spin diffusion measurements: Spin echoes in the presence of a time-dependent field gradient. J Chem Phys 1965;42(1):288–292.
18. Le Bihan D, Breton E, Lallemand D, Grenier P, Cabanis E, Laval-Jeantet M. MR imaging of intravoxel incoherent motions: Application to diffusion and perfusion in neurologic disorders. Radiology 1986: 161(2):401–407.
19. Basser PJ, Mattiello J, LeBihan D. MR diffusion tensor spectroscopy and imaging. Biophys J 1994;66:259–267.
20. Moseley ME, Kucharczyk J, Mintorovitch J, Cohen Y, Kurhanewicz J, Derugin N, Asgari H, Norman D. Diffusion-weighted MR imaging of acute stroke: Correlation with T2-weighted and magnetic susceptibility-enhanced MR imaging in cats. AJNR Am J Neuroradiol 1990;11(3):423–429.
21. Warach S, Dashe JF, Edelman RR. Clinical outcome in ischemic stroke predicted by early diffusion-weighted and perfusion magnetic resonance imaging: A preliminary analysis. J Cereb Blood Flow Metab 1996;16(1):53–59.
22. Witwer BP, Moftakhar R, Hasan KM, Deshmukh P, Haughton V, Field A, Arfanakis K, Noyes J, Moritz CH, Meyerand ME, Rowley HA, Alexander AL, Badie B. Diffusion-tensor imaging of white matter tracts in patients with cerebral neoplasm. J Neurosurg 2002;97(3):568–575.
23. Dumoulin CL, Hart HR. Magnetic resonance angiography. Radiology 1986;161:717–720.
24. Keller PJ, Drayer BP, Fram EK, Williams KD, Dumoulin CL, Souza SP. MR angiography with two-dimensional acquisition and three-dimensional display. Radiology 1989;173:527–532.
25. Prince MR, Yucel EK, Kaufman JA, Harrison DC, Geller SC. Dynamic gadolinium-enhanced three-dimensional abdominal MR arteriography. J Magn Reson Imaging 1993;3: 877–881.
26. Gibbons RJ, Araoz PA. The year in cardiac imaging. J Am Coll Cardiol 2005;46(3):542–551.
27. Association AH, ed. 2005: Heart Disease and Stroke Statistics—2005 Update. Dallas, TX: A.H. Association; 2005.
28. Carr JC, Simonetti O, Bundy J, Li D, Pereles S, Finn JP. Cine MR angiography of the heart with segmented true fast imaging with steady-state precession. Radiology 2001;219(3):828–834.
29. Kim RJ, Wu E, Rafael A, Chen EL, Parker MA, Simonetti O, Klocke FJ, Bonow RO, Judd RM. The use of contrast-enhanced magnetic resonance imaging to identify reversible myocardial dysfunction. N Engl J Med 2000;343(20):1445–1453.
30. Ray T. Magnetic resonance imaging in the assessment of coronary artery disease. Curr Atheroscler Rep 2005;7(2): 108–114.
31. Rickers C, Kraitchman D, Fischer G, Kramer HH, Wilke N, Jerosch-Herold M, et al. Cardiovascular interventional MR imaging: A new road for therapy and repair in the heart. Magn Reson Imaging Clin N Am 2005;13(3):465–479.
32. Bouchiat M, Carver T, Varnum C. Nuclear polarization in 3He gas induced by optical pumping and dipolar exchange. Phys Rev Lett 1960;5:373–375.
33. Albert MS, Cates GD, Driehuys B, et al. Biological magnetic resonance imaging using laser-polarized 129Xe. Nature 1994;370:199–201.
34. van Beek E, Wild J, Kauczor H, Schreiber W, Mugler J, Lange E. Functional MRI of the lung using hyperpolarized 3-helium gas. J Mag Reson Imag 2004;20:540–554.
35. Samee S, Altes T, Powers P, et al. Imaging the lungs in asthmatic patients by using hyperpolarized helium-3 magnetic resonance: Assessment of response to methacholine and exercise challenge. J Allergy Clin Immunol 2003;111: 1205–1211.
36. Salerno M, Lange E, Altes T, Truwit J, Brookeman J, Mugler J. Emphysema: Hyperpolarized helium 3 diffusion MR imaging of the lungs compared with spirometric indexes—initial experience. Radiology 2002;222:252–260.
37. Altes TA, de Lange EE. Applications of hyperpolarized helium-3 gas magnetic resonance imaging in pediatric lung disease. Top Magn Reson Imaging 2003;14:231–236.
38. Moller HE, Chawla MS, Chen XJ, et al. Magnetic resonance angiography with hyperpolarized 129Xe dissolved in a lipid emulsion. Magn Reson Med 1999;41:1058–1064.
39. Callot V, Canet E, Brochot J, et al. MR perfusion imaging using encapsulated laser-polarized 3He. Magn Reson Med 2001;46:535–540.
40. Goldman M, Johannesson H, Axelsson O, Karlsson M. Hyperpolarization of 13C through order transfer from parahydrogen: A new contrast agent for MRI. Magn Reson Imag 2005;23:153–157.
41. Parra-Robles J, Cross AR, Santyr GE. Passive shimming of the fringe field of a superconducting magnet for ultra-low

field hyperpolarized noble gas MRI. J Magn Reson 2005;174: 116–124.

42. Kvistad KA, et al. Breast lesions: Evaluation with dynamic contrast-enhanced T1-weighted MR imaging and with T2*-weighted first-pass perfusion MR imaging. Radiology 2000; 216:545–553.

43. Kuhl CK, Schild HH. Dynamic interpretation of MRI of the breast. JMRI 2000;12:965–974.

MAGNETOCARDIOGRAPHY. See BIOMAGNETISM.

MANOMETRY, ANORECTAL. See ANORECTAL MANOMETRY.

MANOMETRY, ESOPHAGEAL. See ESOPHAGEAL MANOMETRY.

MAMMOGRAPHY

PEI-JAN PAUL LIN
Beth Israel Deaconess
Medical Center
Boston, Massachusetts

INTRODUCTION

It has been shown that reduced breast cancer mortality in the past decade can be attributed to the high sensitivity of screening mammography in detecting nonpalpable lesions (1,2). There has been a wide variety of equipment and imaging modalities employed in the breast cancer detection and imaging; ranging from ultrasound imager, X-ray mammography, computed tomography scanner (CT), to magnetic resonance imager (MRI). Additionally, thermography, light diaphanography, electron radiography, and microwave radiometry have also been utilized, experimentally, to detect breast cancer without much success. Brief explanations of these imaging modalities have been described in the first edition of this Encyclopedia, NCRP Report No.85, and in a review article by Jones (3–5). Among those modalities; ultrasound, X-ray mammography, CT, and MRI, X-ray mammography is the most practical and relatively inexpensive, and is the only main stream of equipment available for breast cancer detection. At present, the ultrasound imager is often employed as an adjunct to the X-ray mammography and is not the primary screening imaging device. However, when mammography is mentioned, it is normally meant to say "X-ray mammography". For these reasons, this article will devote all of its effort to X-ray mammography.

THE MAMMOGRAPHY QUALITY STANDARDS ACT OF 1992 (MQSA), (PUBLIC LAW 102–539)

Breast cancer was a major public health issue in the early 1990s; U.S. Congress enacted MQSA "to ensure that all women have access to quality mammography for the detection of breast cancer in its earliest, most treatable stages". Thus, it is required by law that facilities providing mammography services be properly accredited and be certified by the U.S. Food and Drug Administration (FDA). "Accreditation and Certification" of mammography facilities are beyond the scope of this article, and interested readers are requested to refer to the FDAs WEB Site (6), and the American College of Radiology (ACR) WEB Site (7).

X-RAY MAMMOGRAPHY

Conventional X-ray equipment was initially employed for breast cancer imaging with industrial (thick) emulsion film, or portal imaging film for use in radiation therapy as the image receptor in order to visualize the small microcalcifications, prior to, as late as 1970s. The breast entrance dose of this imaging process exceeded well over 85 mGy per film ($\sim$10 R per film) and the radiographic techniques were typically in the range of 45–55 kVp, and $\sim$1000 mAS with a radiation beam quality of half-value layer (HVL) = 1.0–1.5 mm of aluminum (mmAl) (8). The X-ray beam spectrum produced by a conventional X-ray tube, equipped with tungsten anode, is not necessarily optimized for breast cancer detection. The mammography images obtained in this manner had the desired spatial resolution ($\sim$20 lp·mm^{-1}), but had a less than desirable radiographic contrast. This combination of "high entrance dose" with "low radiographic contrast" was not an acceptable approach. The radiology community was searching for a new breast imaging solution. In the 1970s, xeromammography imaging plates provided the much needed improvement in image quality and lowered the breast entrance dose by a factor of two thirds to one half compared to using the thick emulsion industrial type film (9).

Due to its unique patient positioning of breast imaging, the geometrical arrangement of dedicated mammography units should be pointed out. As shown in Fig. 1, with the

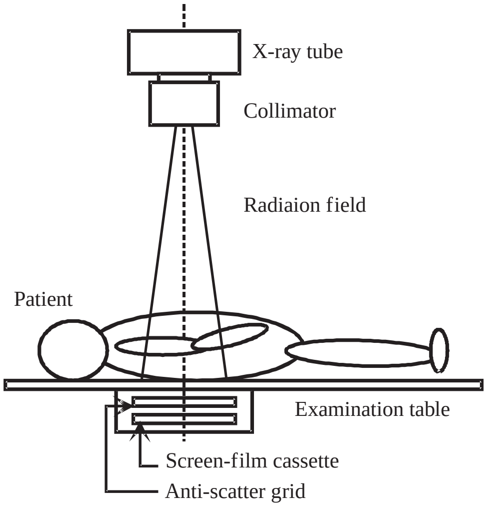
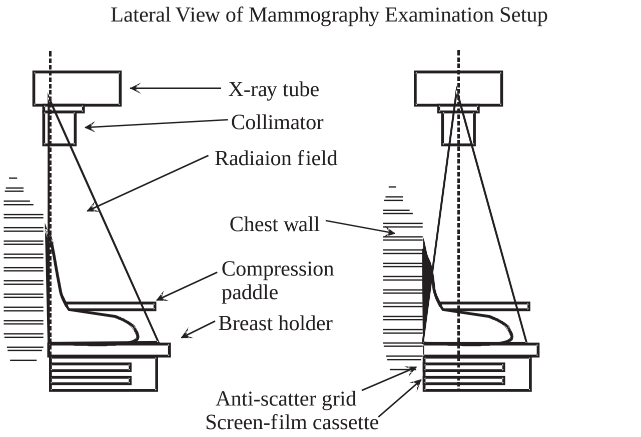


Figure 1. Geometric arrangement of conventional radiography. The X-ray system, the anatomy of interest, and the screen-film cassette are centered and aligned for exposure.



(a) Correct alignment of X-ray system with the patient's chest wall included in the field-of-view.

(b) Improper setup. The "black" triangular shaped area is not in the field-of-view.

Figure 2. Geometry of dedicated mammography systems. On the left side of (a), the geometrical arrangement of a dedicated mammography unit is correctly setup where the center of the radiation beam is aligned to the chest wall of the patient with the compression cone pressing down on the breast being imaged. On the right side (b), improper setup for mammography examination is evident. The "black" triangular area is not included in the image captured by the image receptor, potentially missing the suspected cancer site.

conventional radiography system, the X-ray tube is setup in the center of the anatomy of interest. The radiation field is a projection of diverging X rays restricted by the collimator, centered to the area of interest. The geometry of a dedicated mammography system on the other hand is offset so as to maximize inclusion of breast tissue close to the chest wall (see Fig. 2a). If the geometry were setup in the same manner as the conventional radiography as depicted in Fig. 2b, the black triangular area of the breast would not be included in the imaging field possibly missing a suspected cancer site.

THE PHYSICS OF MAMMOGRAPHIC IMAGING

Considering that various tissues in the breast are radiologically similar, but not identical, the task of differentiating the fibrous, ductal, and glandular tissues is extremely difficult. This is due to the fact that (1) water; the main ingredient of human tissue, has a density of $\rho = 1.0\,\mathrm{g{\cdot}cm^{-3}}$ and effective atomic number of $Z_{\mathrm{eff}} = 7.4$–7.6, and (2) fat has a density of $= 0.9\ \mathrm{g{\cdot}cm^{-3}}$ and effective atomic number of $Z_{\mathrm{eff}} = 5.9$–6.5 (10). Thus, various breast tissues with varying degrees of fat and water contents are very similar from radiological point of view.

The development of screen-film mammography was, in reality, paired with the redesigning of dedicated X-ray equipment for breast imaging. Breasts are relatively thin (physical thickness and X-ray attenuation property) compared to other body parts. Radiographs of extremities, for example, are obtained with X-ray tube potential in the range of 50–60 kVp. Breast tissues contain no high attenuation anatomy, such as bones, a lower tube potential (<35 kVp) would be more suitable for breast imaging (11). Use of lower tube potential has a potential benefit of taking advantage of the photoelectric effect in differentiating the subtle differences of breast tissues.

It should be pointed out that there are five basic ways that X-ray photons interact with matter; they are (1) coherent scattering, (2) photoelectric effect, (3) Compton effect, (4) pair production, and (5) photodisintegration (12). Of these five interactions, photoelectric effect and Compton effect predominantly contribute to the image formation in diagnostic radiology. The probabilities of photoelectric effect and the Compton effect can be expressed as following;

$$\text{Photoelectric effect} \sim Z^3/E^3 \tag{1}$$

$$\text{Compton effect} \sim E * Z \tag{2}$$

In equations (1) and (2), E is the photon energy, and Z is the atomic number. Clearly, from Eq. 1, the lower the X-ray photon energy, and the higher the atomic number the probability of the photoelectric effect will be higher. Since the photoelectric effect is proportional to the "cube" of (Z/E), the differential of breast tissues in Z_{eff} is "enhanced" in the image formation. Thus, a better approach to differentiate and image the breast tissues is to employ a lower X-ray tube potential for imaging. While this approach is "good" for imaging, the radiation dose absorbed by the breast would still be a major concern.

Conventional X-ray tubes employ tungsten as the target material and produce a broad X-ray spectrum through the bremsstrahlung process. At X-ray tube potential of 35 kVp and lower, the tube potential used on most dedicated mammography systems, the tungsten target X-ray tubes would produce broad energy spectrum that contribute less to the image formation and more to the radiation dose. Tungsten, rhodium, and molybdenum are ideal for use in X-ray production due to their relatively high melting points than other metallic elements. Typically, dedicated mammography equipment is equipped with molybdenum target X-ray tube with beryllium window (port), and 30-μm thick molybdenum filter. The X rays, generated at 25–30 kVp tube potential, produced by the molybdenum target X-ray tube contain a large fraction of characteristic X rays at energies $\sim$17–20 keV (13,14). It is, therefore, quite natural to optimize and utilize these characteristic X rays for image formation and, consequently, patient exposure reduction at the same time. Most dedicated mammography equipment employ a combination of molybdenum target with aluminum, molybdenum, or rhodium filters. The characteristic X rays generated at the molybdenum target have X-ray photon energies of 17.4 and 19.6 keV, just below the K-absorption edge of 20 keV, thus are quite transparent to the molybdenum filter. This is illustrated in Fig. 3. On the left of Fig. 3a, is a schematic drawing of the X-ray spectrum generated at 30 kVp with a molybdenum target and aluminum filter. In the middle of Fig. 3b, is the same system with 30 μm thick molybdenum filter. Notice that the K-absorption edge curve (dashed curve) of molybdenum shows that the X-ray photons with energies just above

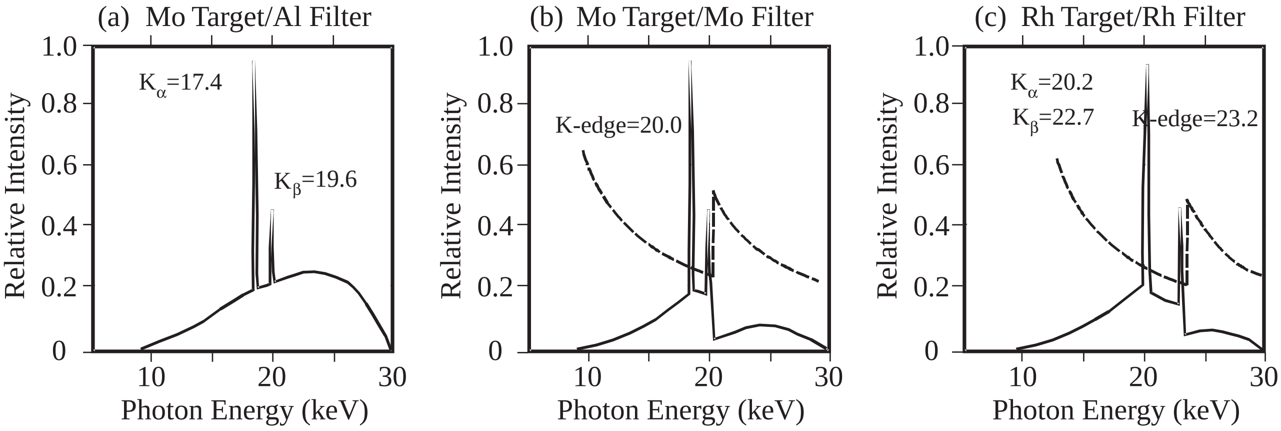


Figure 3. The spectrum of X rays generated with molybdenum and rhodium. On the far left (a) is the X-ray spectrum generated by a molybdenum target at 30 kVp and filtered with aluminum filter. The characteristic X rays K_α (17.4 keV), and K_β (19.6 keV) appear as the two peaks in the graph. In the middle (b) is the X-ray spectrum generated by the same X-ray system in part a, but is filtered with 30 μm thick molybdenum. The K-absorption edge curve is shown in dashed line. On the far right (c) is an X-ray spectrum generated by a rhodium target at 30 kVp and filtered with 20 μm thick rhodium. The general shapes of middle figure and far right figure are similar with differences in the peak energies of K-characteristic X rays ($K_\alpha = 20.2$ keV, and $K_\beta = 22.7$ keV), and the K-absorption edge energy (23.2 keV).

20 keV would be preferentially absorbed than those just below 20 keV. Similarly, the same can be said for the rhodium target with rhodium filter as shown in Fig. 3c. Figure 3b and c are graphically speaking quite similar. It is noteworthy to point out that a careful study of Fig. 3b, and c will reveal that X-ray beams generated from rhodium target X-ray tube "must not" be filtered with molybdenum! The intensity of rhodium K-characteristic X rays (K_α, and K_β) would be in the energy range where the molybdenum K absorption is high. Taking advantage of the spectral information, in 1991,GE introduced the Senographe DMR unit equipped with a dual track and dual filter (molybdenum and rhodium) X-ray tube for mammography applications. Some of the physical characteristics and the spectral energy data of molybdenum, rhodium, and tungsten are summarized in Table 1.

For illustration purposes and descriptions of imaging components, following this paragraph, the screen-film mammography (SFM) system manufactured by GE, the Senogaphe DMR mammography unit, is shown in Fig. 4. The photograph in Fig. 4 represents the overall external and mechanical design of a typical "dedicated" X-ray mammography unit. It represents "the overall" design with respect to the tilting gantry with the X-ray tube housing at the top and the image receptor at the bottom. And, the gantry is attached to a column (or stand), which houses the elevation mechanism of entire gantry.

Table 1. K-Characteristic X Rays and K-Absorption Edge of Molybdenum, Rhodium, and Tungsten

	Molybdenum	Rhodium	Tungsten
Atomic Number	42	45	74
K_α[a]	17.4	20.2	59.3
K_β[a]	19.6	22.7	69.1
K-edge[a]	20.0	23.7	69.5

[a] Energy in keV.

THE SCREEN-FILM MAMMOGRAPHY

The screen-film mammography employs the same basic image receptor system as conventional radiographic imaging. While conventional radiography employs a double-emulsion film sandwiched between two intensifying \ screens yielding a spatial resolution of (up to) 8 lp·mm^{-1} for a detailed screen (15), a typical SFM image receptor consists of a single-emulsion film and a single thin rare-earth phosphor intensifying screen yielding a spatial resolution of ~20 lp·mm^{-1} (16).

In order to optimize the efficiency of the SFM, manufacturers including Agfa, Kodak, Konica, and Fuji, have produced matching pairs of the intensifying screen and film specifically for use with mammography. And, to further improve the sensitometric characteristics of the screen-film, the processing chemistry, particularly the developer, have also been carefully prescribed along with its development conditions. Note that the sensitometric characteristics of screen-film system refers to the "photographic effect or blackening effect" of the screen-film system in response to the X-ray absorption (17). An example of this matching pair of intensifying screen and film is depicted in Fig. 5, the emission and absorption

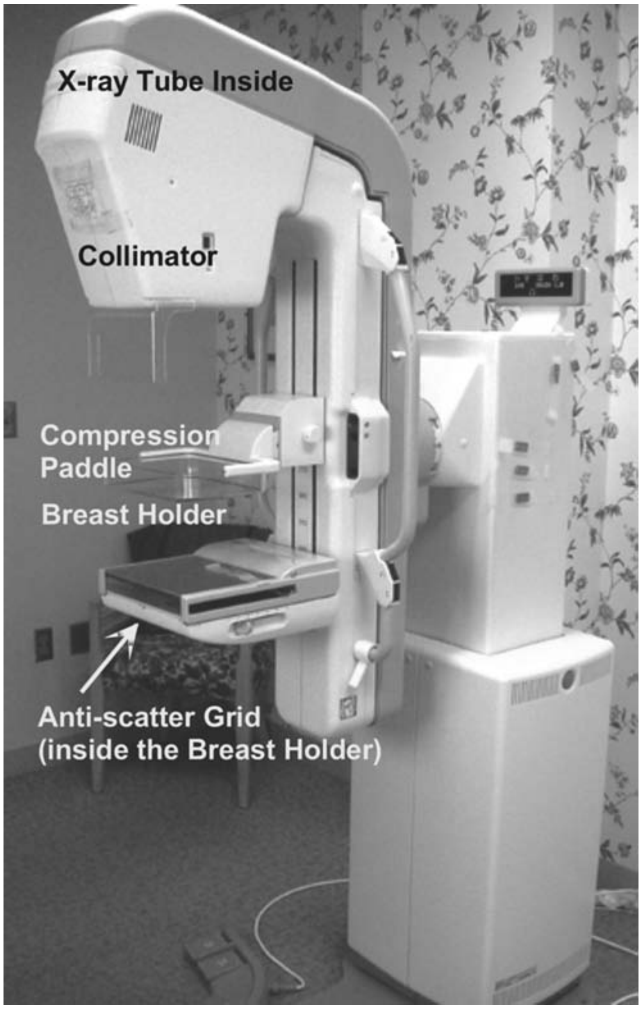


Figure 4. A typical dedicated mammography unit; GE Senographe DMR. (Courtesy of GE Healthcare.) The overall mechanical design of a typical dedicated X-ray mammography showing the tilting gantry with the X-ray tube housing at the top and the image receptor at the bottom. The gantry (or the elongated C arm) is normally attached to a column or a stand in which the elevation mechanism of entire gantry is housed.

characteristics of intensifying screen and film employed in SFM. Note that the emission of 550 nm wavelength light from AD Mammo Screens is matched by the absorption spectra of the AD-M Film.

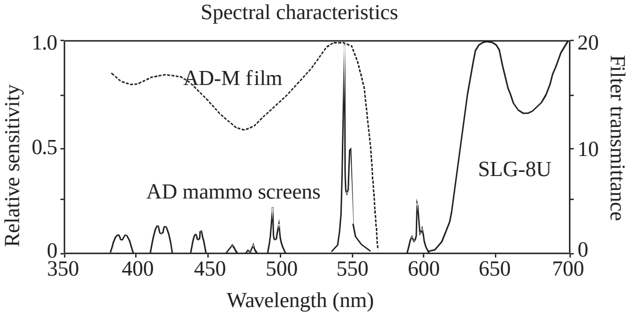


Figure 5. Spectral characteristics of screen-film mammography system. The AD Mammo screens are green-emitting and the AD-M film is orthochromatic. The figure shows the light-emitting spectrum of the AD Mammo Screens, the spectral sensitivity curve of the AD-M film, and the transmittance spectrum of SLG-8U safety light filter. (Courtesy of Fujifilm Medical Systems USA, Inc.)

In order to accommodate the varying size of breasts, two imaging cassette sizes are available with dedicated mammography equipment, they are 18 × 24 cm (8 × 10 in.), and 24 × 30 cm (10 × 12 in.). The film is loaded on the topside of the cassette with the emulsion side facing the intensifying screen. Figure 6 is a schematic drawing showing the cross-section of a typical screen-film mammography cassette (the single emulsion film, single intensifying screen) with enlarged views of the film (top, right), and the screen (bottom, right). The "sponge", in the cassette, is employed to assure good screen-film contact.

THE ANTISCATTER GRID

The antiscatter grid is placed between the breast holder, and the cassette slot, refer to Fig. 4. Scattered radiation is one of the main causes of degrading the image quality in X-ray imaging. The antiscatter grid is employed to minimize the scattered radiation from reaching the screen-film system while allowing the primary radiation to pass through. Although, mammography examinations are typically conducted with X-ray potentials <30 kVp, would still require a moving (reciprocating) antiscatter grid to cleanup the scattered radiation. Typically, the antiscatter grid used in mammography has a grid ratio of 5 : 1, or 4 : 1. While the exposure time in X-ray mammography imaging is relatively long (∼1 s), the speed of the moving grid must be

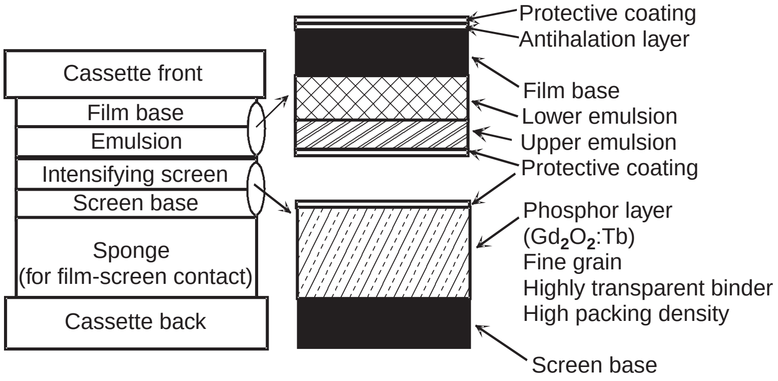


Figure 6. The cross-sectional view of screen-film cassette. (Courtesy of Fujifilm Medical Systems USA, Inc.)

carefully adjusted to avoid artifact associated with grids. For the antiscatter grids, the grid line rate must be sufficiently high, or the attenuation material structure must be so designed to avoid being imaged as grid artifacts on the mammograms. The honeycomb design antiscatter grid; Cellular (HTC) Grid utilized in LoRad mammography systems and marketed by Hologic is well known for its high transmission of useful primary radiation with increased absorption of scattered radiation (18).

THE COMPUTED RADIOGRAPHY FOR MAMMOGRAPHY

Computed radiography was introduced to radiology in 1981 and the "digitization" of routine radiographic examinations had started in earnest. The CR image plate housed in cassette replaced, directly, the screen-film cassette in routine radiography applications. This direct replacement design required no mechanical modifications on the existing radiographic equipment. With the introduction of CR, a whole new set of technology including the optical CR readers, image processing software programs, image display subsystems, and so on, made the filmless radiology department within an achievable reality (19).

The difference between the routine CR and the CR for mammography (CR-M) is largely due to the optical response of the CR phosphor; thus, the radiation dose required to reduce the image noise, and the spatial resolution capability for mammography applications. In 2001, Fuji introduced the FCR 5000MA, which was used in the America College of Radiology Imaging Network (ACRIN); Digital versus Screen-Film Mammography (DMIST) study (20,21). While the study had already been concluded, the official results are in the final preparation stage, and have not been released yet. The FCR 5000MA differentiated itself from the previous generation of CR products by utilizing a 50 μm pixel spot size for the laser and introduced dual side IP reading to the market.

In January 2004, Fuji introduced the Profect CS, or "ClearView-CS" in the United States. Clear View CS is pending FDA approval and not yet commercially available in the United States. Fuji recently finished the Premarketing Approval (PMA) application to FDA, and the approval is anticipated sometime during the second half of 2005 (22). Note, however, that the 100 μm pixel spot size digital mammography system has been in use for the past few years in Europe, Australia, and Japan. The mechanism in which "how" the IP and the CR reader work together to produce an image is beyond the scope of this article and readers are suggested to turn to publications available (23,24), or corresponding articles in this Encyclopedia.

The obvious advantage of CR-M, and the full field digital mammography (FFDM) systems, is its wide dynamic range and its linear response to radiation. In screen-film mammography, over-or underexposure was one of the main causes of "repeat" examinations. Such exposure related repeats can be minimized due to the wide latitude in exposure acceptable for imaging. Furthermore, the images are no longer "fixed" or "limited"; the subject contrast and the overall image brightness can be adjusted for image interpretation by adjusting the display window "level", and "width". The display "window level" corresponds to the film optical density in SFM, and the "widow width" corresponds to the contrast. In digital imaging systems, both window level and width may be adjusted to optimize the image rendered on the monitor. Figure 7 shows the response differences of these two image receptor systems to the radiation exposure. The exposure range for the CRs "Image Plate" is wider than that of a typical screen-film combination by an order of 2 to 3 in magnitude.

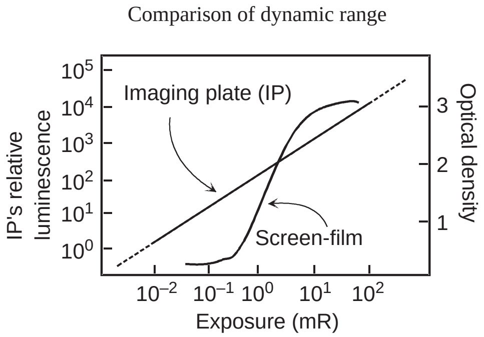


Figure 7. Comparison of dynamic range of CR and screen-film image receptors. (Courtesy of Fuzifilm Medical Systems, USA, Inc.) The image plate (IP) has wider linear response to radiation extends over four decades of dose and eliminates the inherent limitations of the "toe" and "shoulder" portions of the film's H&D Curve. (Courtesy of Fujifilm Medical Systems USA, Inc.)

THE STEREOTACTIC BREAST BIOPSY MAMMOGRAPHY

Breast biopsy is often required when an area of suspicious malignancy site is revealed after a mammography is obtained. In order to accurately extract the specimen, a stereo mammogram may be obtained so that the biopsy needle can be accurately inserted to the site where it is suspected of malignancy; core biopsy (25). Most dedicated mammography equipment are designed to perform routine (screen-film, or FFDM) mammography, and with an attachment to perform stereotactic mammography and core biopsy in an "up right" posture; either the patient is standing or sitting on a chair.

From a pair of stereo images separated by 30° (±15° from the centerline), using the triangulation technique, the three-dimensional (3D) coordinates of the suspicious site can be calculated within an accuracy of 1 mm (25), see the schematic diagram of stereotactic imaging geometry in Fig. 8. It is essential that the patient remain still before and after the acquisition of the stereo images. The core biopsy can be localized accurately only if the patient positioning remain the same.

Stereotactic mammography requires that the breast being examined is compressed (just as in routine mammography), but the patient should remain standing still while the mammogram is being processed. The processing time alone takes at least 90 s with typical automatic film processor. Thereafter, the best location of the biopsy needle entrant site, and the coordinates of the sampling must be determined. The prolonged time represents substantial

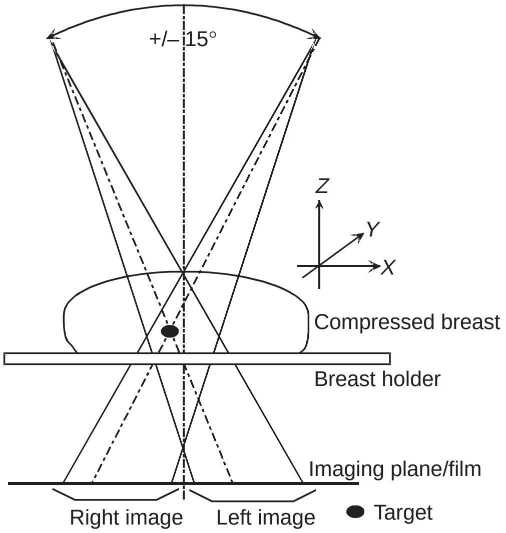


Figure 8. Schematic drawing of geometrical setup for stereotactic mammography. From the stereo images, with known geometry and the *X–Y* coordinates, the depth in *Z* direction is calculated via triangulation.

discomfort on the part of the patient not to mention the anxiety of fear for the possibility of being diagnosed and confirmed as having breast cancer.

DIGITAL SPOT MAMMOGRAPHY

Before the introduction of FFDM, in order to ease the discomfort of undergoing the biopsy process, a small detector with imaging area of 5 × 5 cm had been developed by LoRad, for example, so that the stereo images of the breast can be displayed immediately after exposures are made. This is one of the successful biomedical applications that have resulted from collaborative technology transfer programs between the National Aeronautics and Space Administration (NASA), the National Cancer Institute (NCI), and the U. S. Department of Health and Human Services Office on Women's Health (OWH) (26). Since the imaging area is only 5 × 5 cm, it is referred to as Digital Spot Mammography. The biopsy coordinate is then calculated via the built-in software program with a shorter overall examination time.

In addition, dedicated breast biopsy systems have been developed where the patient would be lying on her stomach (recumbent). The breast under examination is positioned through a hole in the examination table and aligned with the X-ray equipment and the biopsy needle gun. The recumbent core biopsy mammography unit manufactured by LoRad is depicted in Figs. 9 and 10. LoRad is now one of the brand names sold under Hologic. A similar recumbent core biopsy unit, called Mammo Test Biopsy System, is available from Fisher Imaging.

FULL FIELD DIGITAL MAMMOGRAPHY

The CR-M is a direct replacement of the screen-film cassette. In other words, the IP cassette replaced the screen-film cassette. Therefore, no major mammography equipment modification would be necessary. In DR on the other hand, just as in conventional radiography and fluoroscopy applications, the image receptor system may be built into the image receptor compartment as an integral part of the imaging assembly (27,28). Manufacturers had attempted to physically fit the DR detector assembly, or more accurately; the Flat Panel (FP) detector assembly, into the space occupied by the SFM cassette. To date, no commercial product of FP detector assembly has been fitted as a *direct physical replacement* of the SFM cassette is available (29). In other words, due to technical difficulties, there is no FP detector assembly that is sufficiently compact to fit into the space designed to accommodate the SFM cassette. The image acquired with FP detector is transmitted to and processed by the image processing chain thereafter.

While there are a hand full of FFDM systems either under development or manufactured for clinical applications, three FFDM units are currently available in the United States, and are described here (27). The fact that

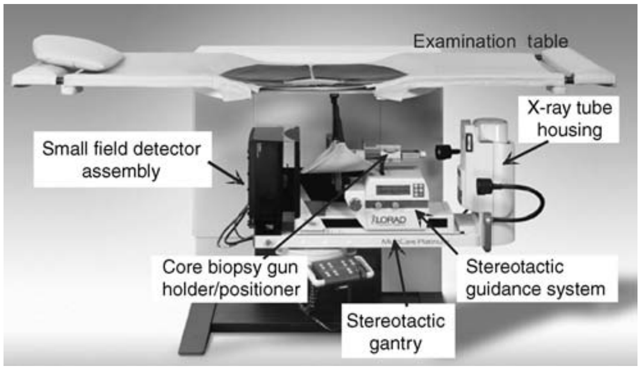

Examination table
X-ray tube housing
Small field detector assembly
Core biopsy gun holder/positioner
Stereotactic guidance system
Stereotactic gantry

Figure 9. The recumbent stereotactic core biopsy system. A recumbent design provides a stable patient positioning to acquire a pair of stereo images for accurate localization of biopsy site while the patient lays flat on her stomach with comfort during the entire examination. (Courtesy of Hologic.)

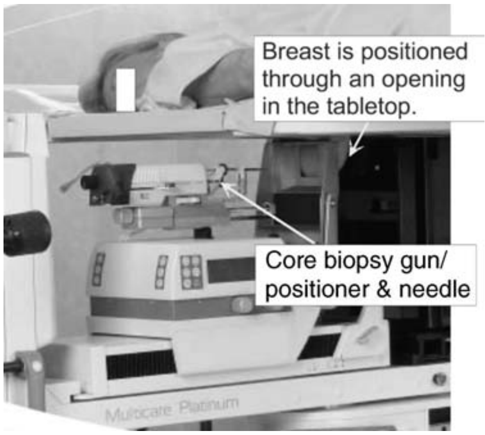

Breast is positioned through an opening in the tabletop.
Core biopsy gun/ positioner & needle

Figure 10. The zoomed view of the recumbent stereotactic core biopsy system. In the zoomed view, the breast under the examination is positioned through the opening in the tabletop. (Courtesy of Hologic.)

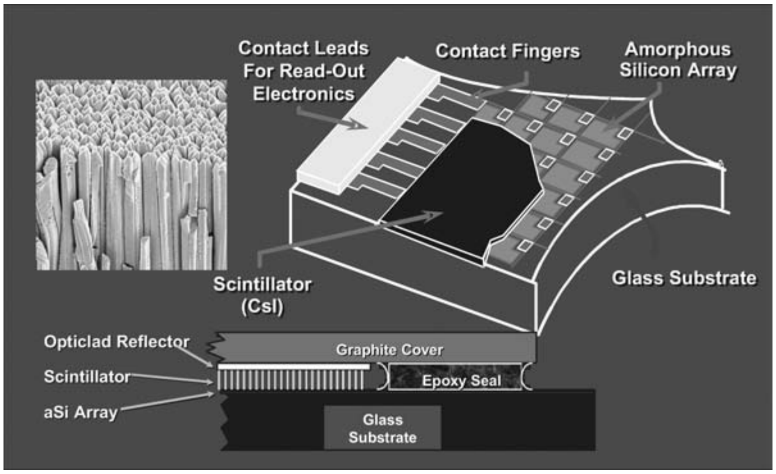


Figure 11. Schematic drawing of an amorphous silicon digital detector. (Courtesy of GE Healthcare.)

these three systems are designed to cover an imaging field of at least (~) 18 × 24 cm, the smaller cassette size of SFM, the name; FFDM was assigned. This is to signify the full size imaging filed coverage as opposed to the small field detector of LoRad's Digital Spot Mammography, or Fischer Imaging's Mammo Test Biopsy System.

There is a trend that mammography equipment is being transformed to "digital" format. One reason of the slow pace of this transformation is the large initial capital expenditure of FFDM. Both, analog and digital formats are expected to coexist for many more years to come. The impact of Fuji's CR-M cannot be ignored due to the fact that it is possible to convert and replace the screen-film cassette directly. Any of the existing SFM units will be able to join the "digital era".

Currently, there are three FFDM systems that have received FDA approval and are sold in the United States. The GE Senogaphe 2000D was approved by FDA on January 28, 2000 and "accredited" by ACR in February, 2003. Fischer Senoscan received its FDA approval on September 25, 2001 with subsequent ACR accreditation in August, 2003. Hologic/LoRad's Selenia received FDA approval and ACR accreditation, respectively, on October 2, 2002, and in September of 2003 (27).

GEs FFDM: SENOGRAPHE 2000D, AND SENOGRAPHE DS

Figure 11 depicts the FP detector developed by General Electric Medical Systems, and the photo inset (top left) shows the Cesium–Iodide (CsI) scintillator crystals. The incident X rays are absorbed by the CsI crystals, which in turn, convert the X-ray photon energy to light. The CsI crystals are deposited in columnar shape so as to minimize the scatter, and optical diffusion of scintillation lights from one column to the other. The underlying photodiode–transistor amorphous Silicon (a-Si) arrays is connected to control data lines, then, converts the light to electrical signals for further processing (30).

GEs FP detector is an indirect-conversion digital detector system. The detective quantum efficiency (DQE) of this FP detector system has been shown to be ~60% at Zero Frequency (31,32). Essentially, being the first FFDM system introduced to the commercial market, GEs Senographe 2000D ushered in the digital mammography era to the radiology community. Depicted in Fig. 12 is the second-generation FFDM unit, Senographe DS, with the X-ray gantry set to + 15° with the stereotactic biopsy needle assembly attached. Both Senographe systems (2000D and DS) are equipped with CsI scintillator on an a-Si photodiode–transistor arrays with pixel size of 100 μm and an image matrix size of 1920 × 2304, covering a field of view 19 × 23 cm.

FISCHER IMAGING'S SENOSCAN FFDM

On the other hand, Fischer Imaging Corporation's answer to the FFDM is the Senoscan FFDM unit. The unique feature of Senoscan unit is that it employs a slot mechanism for the X-ray filed collimation that is synchronized to the detector as the slot and detector assembly sweeps across the imaging field, see Fig. 13. The detector of

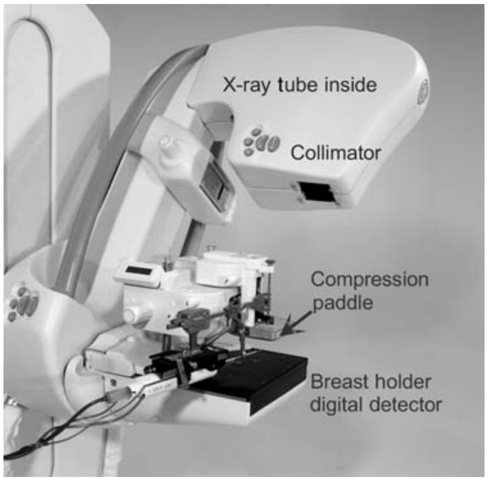


Figure 12. Photograph of GE FFDM unit; Senographe DS with stereotactic attachment installed. The gantry is tilted to +15° and prepped for Stereotactic imaging. Note, this is the FFDM version of similar unit shown in Fig. 1, but with zoomed up view for the stereotactic attachment. (Courtesy of GE Healthcare.)

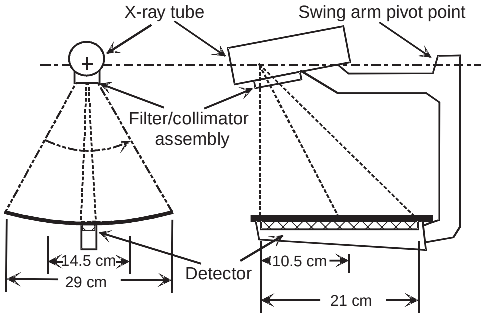


Figure 13. A schematic drawing of the scanning mechanism of Fischer Senoscan FFDM. (Adapted from Fisher Imaging Sanoscan user's manual with permission. Courtesy of Fischer Imaging.)

Senoscan unit consists of a layer of thallium activated CsI scintillator connected to charge-coupled-device (CCD) chips via fiber optics (33). Hence, this is also an indirect-conversion digital detector system.

The use of slot mechanism also means "restricting" the X-ray beams for exposure. It offers an advantage of less scattered radiation for improved image contrast (34). Therefore, no antiscatter grid is necessary with Senoscan unit. The absence of the antiscatter grid compensate for a less efficient use of the available X rays. In addition, a tungsten anode X-ray tube (35) is employed with this unit, since the tungsten anode X-ray tube would produce X rays more efficiently than a molybdenum or rhodium target. Senoscan is equipped with CCD chips having pixel size of (1) 27 μm, covering a field of view 11 × 15 cm under the high resolution mode, and (2) pixel size of 54 μm, covering a field of view 19 × 29 cm for the normal mode. The image matrix size of this system is 4096 × 5625.

HOLOGIC/LORAD'S FFDM; SELENIA

Selenia's image receptor is a 250 μm thick amorphous Selenium (a-Se) photoconductor. Underneath a layer of a-Se photoconductor is a thin-film transistor (TFT) arrays that serves as an active readout mechanism. The TFT arrays are typically deposited onto a glass substrate, which provides a physical support for the entire detector components. Selenia uses a 250 μm thick a-Se photoconductor to capture the X-ray photons impinging on the detector surface without the aid of a scintillator. It is said that a thickness of 250 μm a-Se photoconductor is adequate to stop 95% of the X-ray photons in the mammographic energy range (36). The photon energy is converted to a pair of electron, which is negatively charged, and a positively charged "hole". With bias voltage applied, the signal is read off by the TFT arrays. Thus, this is a direct-conversion digital detector system. The detector is an a-Se photoconductor and TFT arrays with pixel size of 70 μm, covering a field of view 24 × 29 cm, resulting in an image matrix size of 3328 × 4096. A summary of the detector characteristics for these three FFDM systems is given in Table 2, (37).

READING MAMMOGRAPHY IMAGES

Initially, the FFDM images were printed on dry laser printers and read on high luminance view boxes (ACR accreditation required of a luminance of $>3000\,\mathrm{cd{\cdot}m^{-2}}$) (38). All three FFDM systems are equipped with workstations where images are soft-copy read. The workstations are commonly equipped with two 5-megapixel high resolution monitors. With the soft-copy reading, an assortment of image manipulation are possible including, but not limited to, basic window width and window level adjustments, zooming, image reversal, and so on for viewing the details of the pathology. More importantly, the impact of image processing in transforming the acquired raw data set to the image displayed for soft-copy reading should be recognized (39). For example, as can be seen in Fig. 14, substantial differences not only in the brightness and contrast of the image but also in the impression of pathology in the images can be noticed. While all three images are acceptable and diagnostic, the two enhanced images are easier to recognize various details.

SUMMARY

A brief history of mammography was described as the introduction to the X-ray mammography. The SFM continues to play its major role in breast cancer detection while FFDM is gaining its installed base. Upon the anticipated FDA approval, the CR-M is poised for introduction for clinical applications in the second-half of 2005. However, all three modalities employed in breast cancer detection, namely; the SFM, FFDM, and the CR-M are expected to

Table 2. Summary of Imaging Characteristics of FFDM Units

Manufacturer	Model Name	Scintillator	Detector	Pixel Size, μm	Image Matrix Size	Imaging Size, $L \times W$ cm
GE	Senographe 2000D, DS	CsI (Tl)[a]	TFT	100	1920 × 2304	19 × 23
Fisher Imaging	Senoscan	CsI (Tl)	CCD	24/48	4096 × 5625	22 × 30
Hologic/LoRad	Selenia	a-Se[b]	TFT	70	3328 × 4096	25 × 29
	DSM	CsI (Tl)	CCD	48	1025 × 1024	5 × 5
Fuji	CR-M	BaFBr(Eu)[c]	Computed	50	1770 × 2370	18 × 24
			radiography		2364 × 2964	24 × 30

[a] CsI(Tl) = Talium Activated Cesium Iodide.
[b] a-Se = Amorphous Selenium.
[c] BaFBr(Eu) = Barium Fluorobromide with Europium.

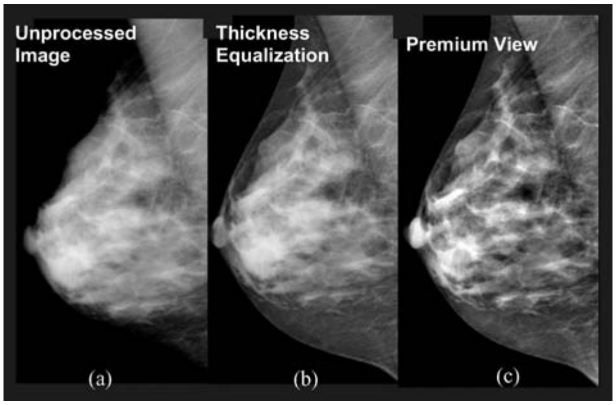


Figure 14. Comparison of "processed" images. The unprocessed image (a) can be dramatically improve its appearance with "Thickness Equalzation" processing (b), or the "Premium View" processing (c). The impact of image processing is quite evident. (Courtesy of GE Healthcare.)

play their roles in contributing to reduction of breast cancer deaths through the early detection of mammography screening.

Display of FFDM images and its workstations are important components of the total picture of FFDM (40,41). However, monitors and workstations associated with diagnostic radiology imaging including mammography are in a domain of their own and no attempt is made to include these subjects in this article. Such subject matters belong to digital image processing and display. Finally, archiving of the acquired digital images is also a very important aspect of digitized diagnostic images in the overall operation of radiology. However, this subject is better handled in Picture Archiving and Communication Systems (PACS), and readers are refered to articles under PACS for additional information.

ACKNOWLEDGMENT

The information and materials presented here had been provided by numerous numbers of representatives and scientists from respective manufacturers mentioned in the main text. The author would like to express his thanks and sincere appreciation by listing the manufacturers here, (in alphabetical order) in lieu of listing the individuals who had provided their supports: General Electric Healthcare, Fischer Imaging Corporation, Fujifilm Medical Systems USA, Inc., and Hologic Inc. (LoRad).

BIBLIOGRAPHY

1. Nystrom L, et al. Breast cancer screening with mammography: Overview of Swedish randomised trial, Lancet 1993; 341:973–978.
2. Hendrick RE, et al. Benefit of screening mammography in women aged 40–49: A new meta-analysis of randomized controlled trial. J Nat Cancer Inst Monograph 1997; (22):87–92.
3. Mammography—A User's Guide, Washington, DC: National Council on Radiation Protection and Measurements, 1986. NCRP Report No. 85.
4. Zermenhof RA. Mammography, Encyclopedia of Medical Devices and Instrumentation, Vol 3. New York: John Wiley & Sons; 1988.
5. Jones CH. Methods of breast imaging. Phys Med Biol 1982;27(4)463–499.
6. FDA Home page. [Online]. Available at http://www.fda.gov/cdrh/mammography/frmamcom2.html.
7. ACR Home page. [Online]. Available at http://www.acr.org/s_acr/sec.asp?CID=589&DID=14253.
8. Speiser RC, Zanrosso EM, Jeromin LS. Dose comparisons for mammographic systems. Med Phys 1986;13(5):667–673.
9. Boag JW. Xeroradiography. Phys Med Biol 1973;18:3.
10. Wolbarst AB. Dependence of attenuation on atomic number and photon energy. Physics of Radiology. Appleton & Lange; 1992. Chapt. 14.
11. Hoeffken W. Soft tissues—mammography. The invisible light applied, the advancing diagnostic evaluation of the skeleton and thorax following Roentgen's discovery. In: Rosenbusch G, Oudekerk M, Ammann E, editors. Radiology in Medical Diagnostics, Evolution of X-ray Applications 1895–1995. Oxford: Blackwell Science Ltd.; 1995. Chapt. 2.
12. Curry III TS, Dowdey JE, Murry Jr. RC. Christensen's Physics of Diagnostic Radiology, 4th ed., Philadelphia: Lea & Febiger; 1990.
13. Fewell TR, Shuping RE. Handbook of Mammographic X-ray Spectra. Rockville, MD: HEW Publication (FDA) 79–8071; 1978.
14. Boone JM, Fewell TR, Jennings RJ. Molybdenum, rhodium, and tungsten anode spectral modeling using interpolating polynomials with application to mammography. Med Phys Dec. 1997:24(12).
15. Sprawls Jr. P. Chapter 18: Blur, Resolution, and Visibility of Detail. Physical Principles of Medical Imaging. Aspen Publishers, Inc.; 1987.
16. Haus AG. Technologic improvements in screen-film mammography. Radiology 1990;174:628–637.
17. Chapter V. Film Response. The Fundamentals of Radiography, 4th ed., Rochester, N.Y.: Eastman Kodak Co.; 1980.
18. Rezentes PS, de Almeida A, Barnes GT. Mammography grid performance. Radiology 1999;210:227.
19. Sonoda M, Takano M, Miyahara J, Kato H. Computed radiography utilizing laser stimulated luminescence. Radiology 1983;148:833–838.
20. Pisano ED, et al. American College of Radiology Imaging Network Digital Mammographic Imaging Screening Trial: Objectives and Methodology. Radiology published online. June 16, 2005.
21. ACRIN Home Page. [Online]. Available at http://www.acrin.org.
22. Communication with National Program Manager, Women's Healthcare Imaging Systems, Fuji Photo Film Ltd.
23. FCR (Fuji Computed Radiography) General Description of Image Processing, Fuji Photo Film Co., Ltd.
24. Tateno Y, Iinuma T, Takano M, editors. Computed Radiography. Tokyo: Springer-Verlag; 1987.
25. Carr JJ, et al. Stereotactic localization of breast lesions: How it works and methods to improve accuracy. RadioGraphics 2001;21:463.
26. Winfield DL. Aerospace technology transfer to breast cancer imaging. Acta Astronau 1997;41(4–10):515–523.
27. Pisano ED, Yaffe MJ. Digital Mammography. Radiology 2005;234:353–362.
28. Pisano ED. Current status of full-field digital mammography. Radiology 2000;214:26.
29. Personal communications with various X-ray industry marketing departments.

30. Vedantham S, et al. Full Breast digital mammography with anamorphous silicon-based flat panel detector: Physical characteristics of a clinical prototype. Med Phys 2000; 27(3):558–567.
31. Albagli D, et al. Performance of advanced a-Si/CsI-based flat panel X-ray detectors for mammography. In: Yaffe MJ, Antonuk LE, editors. Proceedings of SPIE. Medical Imaging 2003: Physics of Medical Imaging, Vol. 5030, June 2003. p 553–563.
32. Shaw J, Albagil D, Wei C-Y. Enhanced a-Si/CsI-based flat panel X-ray detector for mammography. In: Yaffe MJ, Flynn MJ, editors. Proceeding of SPIE. Medical Imaging 2004: Physics of Medical Imaging. Vol. 5368, May 2004. p 370–378.
33. Tesic MM, Picaro MF, Munier B. Full field digital mammography scanner. Eur J Rad 1997;31:2–17.
34. Boone JM, et al. Grid and slot scan scatter reduction in mammography: Comparison by using Monte Carlo techniques. Radiology 2002;222:519–527.
35. Operator Manual, SenoScan: Full Field Digital Mammography System, Fischer Imaging Corporation, Denver, CO, Dec. 2002.
36. Smith AP. Fundamentals of digital mammography: Physics, technology and practical considerations. Radiol Manager 2003, Sep.–Oct.; 25(5):18–24, 26–31.
37. Mahesh M. AAPM/RSNA physics tutorial for residents, Digital mammography: an overview. Radiographics 2004;24: 1747–1760.
38. The ACR Mammography Quality Control Manual. Preston, VA: ACR, 1999.
39. Pisano ED, Yaffe MJ. Digital mammography. Radiology 2005;234:353–362.
40. Hemminger BM, et al. Evaluation of the effect of display luminance on the feature detection of simulated masses in mammograms. Proc SPIE 1997;3036:12.
41. Pisano ED, et al. Radiologists' preferences for digital mammographic display. Radiology 2000;216:820–830.

Further Reading

Mahesh M, AAPM/RSNA physics tutorial for residents, Digital mammography: An overview. *Radiographics* 2004;24:1747–1760.

Smith AP, Hall PA, Marcello DM. Emerging technologies in breast cancer detection. *Radiology Manager* 2004, July.–Aug.; 26(4): 16–24.

Shramchencko N, Blin P, Mathey C, Klausz R. Optimized exposure control in digital mammography. In: Yaffe MJ, Flynn MJ, editors. Proceedings of SPIE. Medical Imaging 2004: Physics of Medical Imaging. May 2004; Vol. 5368, p 445–456.

Curry III TS, Dowdey JE, Murry Jr. RC. Christensen's Physics of Diagnostic Radiology. 4th ed., Philadelphia: Lea & Febiger; 1990. (This is a textbook used in various radiology residency programs across the United States.)

Online References

FDA Home page. [Online]. Available at http://www.fda.gov/cdrh/mammography/frmamcom2.html.

ACR Home page. [Online]. Available at http://www.acr.org/.

ACRIN Home Page. [Online]. Available at http://www.acrin.org.

GE Healthcare Home Page. [Online]. Available at http://www.gehealthcare.com/usen/whc/whcindex.html.

Fischer Imaging Home Page. [Online]. Available at http://www.fischerimaging.com/default/default.asp.

Fujifilm Medical Systems USA Home Page. [Online]. Available at http://www.fujimed.com.

Hologic Home Page. [Online]. Available at http://www.hologic.com/.

MATERIALS, BIOCOMPATIBILITY OF. See BIOCOMPATIBILITY OF MATERIALS.

MATERIALS, PHANTOM, IN RADIOLOGY. See PHANTOM MATERIALS IN RADIOLOGY.

MATERIALS, POLYMERIC. See POLYMERIC MATERIALS.

MATERIALS, POROUS. See POROUS MATERIALS FOR BIOLOGICAL APPLICATIONS.

MEDICAL EDUCATION, COMPUTERS IN

ARIE HASMAN
Maastricht
The Netherlands

INTRODUCTION

The amount of knowledge is increasing rapidly in many disciplines. Medicine is not an exception. Because of scientific research new knowledge comes available at such a pace, that physicians should read 19 articles a day, every day of the week, to keep up to date. Since that is not possible the results of scientific research often are applied clinically only years later. Computers can support physicians in finding relevant recent information. In the next section, the reasons for using computers in medical education are presented. Then the roles computers can play in medical education are reviewed. Since health professionals increasingly use computer systems for their work, they need to know the benefits and limitations of these systems. The discipline of medical informatics is responsible for developing these systems and therefore is discussed. The next sections discuss the use of Internet and electronic patient records. Also, it is discussed why knowledge of information systems is important for health professionals.

THE REASONS FOR USING COMPUTERS IN MEDICAL EDUCATION

The goal of academic medical education is to educate students to become physicians. During the study knowledge, skills and attitudes have to be mastered. Students are taught academic skills like critical reading, they are acquainted with the principles of research methods, and they should know the scientific background of the basic disciplines (like anatomy, molecular cell biology and genetics, endocrinology and metabolism, immunology and inflammation, growth, differentiation and aging) as far as they are related to the study of abnormalities and to diagnosis and therapy. Because of the explosive growth of biomedical knowledge, it is not possible anymore to teach all the currently available knowledge. This does not have to be a problem since part of the knowledge presented to medical students during their formal education may be more or less obsolete by the time they are in their main professional practice. Moreover, it is difficult to teach medicine for the coming era, since most of the future's technology is probably nonexistent today. Students therefore must be taught how to become life-long learners. Computers can support this process.

Computers play an increasing role in the practice of medicine too. No doctor—whether a general practitioner or a specialist in advanced or social care—will be able to escape the confrontation with some form of information processing. The work of health professionals is dominated by information collection, storage, retrieval, and reasoning. Health professionals both use individual patient data and general medical or nursing knowledge. The amount of medical and nursing knowledge increases so quickly that health professionals cannot stay fully up to date. Tools are therefore needed to acquire relevant knowledge at the time it is needed.

Computer systems are installed in many hospital departments and physician offices. Hospital information systems support, for example, financial, administrative, and management functions. Clinical departmental systems are used to collect, store, process, retrieve, and communicate patient information. Clinical support systems are used in function laboratories [for electroencephalogram (EEG), electrocardiogram (ECG), electromyogram (EMG), and spirometry analysis], for imaging [magnetic resonance imaging (MRI), computerized tomography (CT), nuclear medicine, ultrasound], and in clinical laboratories (analysis of electrolytes, etc.). The results of clinical support systems are increasingly stored in so-called electronic patient records, together with the medical history, results of physical examination, and progress notes. Electronic patient records gradually replace the paper patient record. Apart from the fact that electronic paper records are better readable they also support functions (like decision support) paper records cannot provide. Students must learn the benefits and limitations of these kinds of systems.

Decision support systems are used to support clinicians during the diagnostic or therapeutic process and for preventive purposes to prevent either errors of omission (when, eg, the physician does not order a mammography when indicated) or commission (when, eg, the physician prescribes the wrong drug).

Clinical patient data are increasingly stored in the above mentioned electronic patient records (EPRs), from which they can be later retrieved by physicians or nurses who are in need of the data. Also, information systems can retrieve relevant data from the electronic patient record when a suitable interface between system and EPR is available. When a standard vocabulary is used for representing data values in the EPR, decision support systems can interpret these data and remind, alert, or critique the physician or provide access to relevant knowledge, based on patient data available in the EPR. Health professionals should have insight in and knowledge of the principles, concepts, and methods underlying electronic patient records.

Also, patients become active players in the field and increasingly demand access to the EPR.

Patient management increasingly has become the combined task of a group of healthcare workers. Therefore the memory-aid role of the patient record more and more changes into a communication role. Paper records have several limitations in this respect. In addition, since the appearance of the report “To err is human” of the IOM (1) it is apparent that due to miscommunication (among which are problems with reading handwritten information, with incomplete information, etc.) medical errors are made that even may lead to the death of patients. Electronic patient records and order entry systems can reduce the number of errors because they are not only more readable, but because they can also be interfaced with decision support systems when standardized terminology is used.

Decision support can be passive in the sense that the decision support system contains information that has to be searched by the physician. In this case, the healthcare professional takes the initiative to search for information, for example, via PubMed or the Cochrane Library. Decision support can also be active. In this case, the decision support system volunteers advice based on information it can retrieve from the EPR. Decision support systems can either proactively suggest the next step in a diagnostic or treatment process or reactively remind the healthcare professional that a (preventive) procedure was not performed or a step in the protocol was not carried out.

ROLES FOR COMPUTERS IN MEDICAL EDUCATION

What roles can computers play in medical education? In the first place, information systems can be used to manage the learning process. Students can get access to the curriculum via so-called learning environments (e.g., Blackboard or WebCT), can get overviews of their marks, can access computer aided instruction programs, can access PowerPoint presentations of their teachers, and so on. Computers provide access to the internet so that they can search for content knowledge.

Computer-aided instruction can be used to teach students certain subjects. For example, computers are used to simulate regulatory mechanisms occurring in the human body. With a simulation program, students can treat “patients” without risking their patient lives. The simulation is often based on models that present (patho)physiologic processes in the form of mathematical equations. When the models become increasingly accurate they can even be used in patient care. Also, patient management problems can be simulated. In this case, usually no mathematics is involved, but the patient’s signs and symptoms as they develop as a function of time are expressed as text.

There are simulation tools that allow users to evaluate, plan, or redesign hospital departments, or parts of other healthcare systems. Physicians will be confronted with the results of simulations. The model that is used in the simulation has to be checked for validity. Some tools present the modeled processes visually so that physicians or nurses can easily determine the correctness of the model. An example is the modeling of a phlebography service. On the screen, the cubicles and other rooms are displayed and the movements of patients, physicians, and nurses can be followed. In this way, the users can judge whether the model represents the situation of a phlebography service in an adequate way.

Note that computers should not be considered as surrogate teachers controlling students' learning. Computers should enrich the learning environment by expanding the student's control over their self-learning and by providing a better learning environment as a supplement to traditional methods of learning. The effectiveness of CAI has always been a subject of controversy. Studies have claimed both that CAI is superior and that CAI is inferior to traditional methods. The majority of the publications, however, support the notion that CAI is as effective as traditional educational methods (2).

Although decision making is the pre-eminent function of a physician, hardly anywhere is the student confronted with a systematic exposition of procedures of good decision making. The use of computers can facilitate the teaching of these procedures. Computer support offers new possibilities to teach problem solving techniques and analytical methods that are presently learned by the student through practice and the observation of mentors.

MEDICAL INFORMATICS

Education concerning the advantages and limitations of the use of computers for supporting the work of health professionals is the responsibility of medical informatics departments. Medical informatics can also be instrumental in developing computer-aided instruction programmes and simulation packages. Medical informatics is the discipline that deals with the systematic processing of data, information, and knowledge in the domains of medicine and healthcare. The objects of study are the computational and informational aspects of processes and structures in medicine and healthcare (3). Medical informatics is a very broad discipline covering subjects like applied technology (bioinformatics, pattern recognition, algorithms, human interfaces, etc.) and services and products (quality management, knowledge-based systems, electronic patient records, operations–resource management, etc.). Also human and organizational factors (managing change, legal issues, needs assessment, etc.) should be taken into account.

Informatics can be either a systems-oriented discipline in which computer systems, operating systems and programming languages are the object of study or a methods-oriented discipline in which the methods are studied that can be used to create algorithms that solve problems in some application domain. In the case of the methods-oriented approach, a problem is studied and formalized solutions are determined. Medical informatics is an example of this approach: it studies the processing of data, information, and knowledge in medicine and healthcare. Medical informatics focuses on the computational and informational aspects of (patho)physiological processes occurring in the patient, cognitive processes going on in the brain of physicians, and organizational processes that control healthcare systems.

The resulting knowledge can be used to design information systems that can support healthcare professionals. It is clear that healthcare professionals need to have some medical informatics knowledge in order to optimally use information systems. Medical informatics education should therefore be part of the medical curriculum.

Various groups of professionals with quite different backgrounds can be identified who carry out medical informatics tasks ranging from the use of IT to developing information systems. Users of information systems naturally need less medical informatics knowledge than health informatics experts who develop health information systems or support other healthcare workers in designing terminology servers, coding systems, and so on.

There exists a wide range of job opportunities in the field of medical informatics. These jobs require various medical informatics capabilities. In addition to medical informatics, students (and graduate students) with other backgrounds may prefer a job in the field of medical informatics. In order to obtain the relevant capabilities these students have to learn additional subjects depending on their previous education and the type of specialization they want to achieve. These students can be graduates from healthcare related programs or from informatics–computer science programs. Graduates from healthcare related programs possess the relevant medical knowledge, but need to increase their medical informatics knowledge. Graduates with an informatics or computer science background must learn how the healthcare system is organized and how healthcare professionals are working in order to develop systems that are appreciated by healthcare professionals. Medical informatics is therefore taught in different ways (4) depending on the type of students and the type and extent of specialization that they want to achieve.

USE OF THE INTERNET

Much knowledge can be found on the internet. Browsers allow health professionals and patients to access sites containing (references to) medical knowledge. PubMed is an example. It contains references to the medical literature. The web contains a lot of information of which the quality is not always guaranteed. Especially in the medical arena, this is a big disadvantage. The internet has become one of the most widely used communication media. With the availability of Web server software, anyone can set up a Web site and publish any kind of data that is then accessible to all. The problem is therefore no longer finding information, but assessing the credibility of the publisher as well as the relevance and accuracy of a document retrieved from the net. The Health On the Net Code of Conduct (HONcode) has been issued in response to concerns regarding the quality of medical and health information (5). The HONcode sets a universally recognized standard for responsible self-regulation. It defines a set of voluntary rules to make sure that a reader always knows the source and the purpose of the information they are reading. These rules stipulate, for example, that any medical or health advice provided and hosted on a site will only be given by medically trained and qualified professionals unless a clear statement is made that a piece of advice is offered from a nonmedically qualified individual or organization. Another guideline states that support for the Web site should be clearly identified, including the identities of commercial and noncommercial organizations that have contributed funding, services or material for the site. Students searching for information should be introduced to these guidelines.

Searching can be carried out by entering keywords. These keywords can be connected by Boolean operators like AND, OR, and NOT. A user can for example enter: Diuretics and Hypertension to search for documents that discuss the use of diuretics in hypertension. The NLM (National Library of Medicine) uses the Medical Subject Headings (MeSH) vocabulary for indexing most of their databases. Students should be taught how to efficiently search in bibliographic databases using, for example, the MeSH vocabularies.

We speak of e-learning when content is accessible via Web browsers. Some characteristics of e-learning follow: Internet is the distribution channel. Access to the content is possible 24 h/7 days a week. It is learner-centered. The student determines the learning environment, the speed of learning, the subjects to consider, the learning method. A mix of learning methods can be used (blended learning): for example, virtual classroom, simulations, cooperation, communities, and "live" learning.

Virtual learning environments aim to support learning and teaching activities across the internet. Blackboard and WebCT are examples of such environments. These environments offer many possibilities. New or modified educational modules can be announced or teachers can give feedback regarding the way a module is progressing. Also general information about a module can be provided. Staff information can be presented with photo, email address, and so on. Assignments can be posted, and so on. The virtual classroom allows students to communicate online, whereas discussion boards allow asynchronous communication. Also, links to other websites can be provided. The internet is a source of information for patients. They can retrieve diagnostic and therapeutic information from the internet. Increasingly, patients present this information to their care providers. Health professionals must know how to cope with this new situation and must be able to assess the quality of the information.

KNOWLEDGE OF INFORMATION SYSTEMS

Information systems are increasingly used in healthcare. They not only support administrative and financial, but also clinical and logistic processes. Since healthcare workers have to use information systems they should know the possibilities, but also the limitations, of information systems. Since in information systems, for example, data can be easily retrieved, the quality of entered data determines the quality of the results: garbage in, garbage out. In addition, they should have the skills to work with information systems. Information systems relevant for healthcare professionals include hospital information systems, departmental systems, electronic patient record systems, order entry and result reporting systems, and so on. But healthcare workers should also be proficient in the use of productivity tools like word processing systems, bibliographic search systems, and so on.

Logistics is becoming more important these days. Hospitals have to work not only effectively, but also more efficiently, thereby taking the preferences of patients into account. Planning systems can reduce the time that ambulatory patients have to spend in the hospital for undergoing tests, but also the length of stay of hospitalized patients can be reduced by planning both the patients and the needed capacity (6).

It is important for healthcare workers to know what support they can expect from information systems and to know which conditions have to be satisfied in order that information systems can really be of help. Optimal use of information systems therefore does not only depend on acquired skills, but also on the insight in and knowledge of the principles, concepts, and methods behind information systems. This is true for all types of healthcare professionals. When hospitals or physicians consider the purchase of information systems they must be able to specify their requirements so that they will not be confronted with systems that do not perform as expected.

ELECTRONIC PATIENT RECORDS

Physicians store information about their patients in patient records. The patient record frequently is a paper record in which the physician writes his notes. Paper records have several advantages because they are easy to use, easy to carry, and so on. But there are also limitations: they may be difficult to read and are totally passive: the physician records the information with little support (e.g., headings in a form) and therefore the recordings are often incomplete. Not only are the data incomplete, they also contain errors, for example, due to transcription or because the patient gave erroneous information. The readers of patient records can interpret the data in the patient record incorrectly. The fact that data are recorded in chronological order makes the retrieval of facts sometimes difficult: such data are not recorded on standard positions in the record. A study showed that because of the constrained organization of paper records, physicians could not find 10% of the data, although these data were present in the paper record (7). The patient data are usually stored in more than one type of record, because each department uses its own records, each with a different lay-out. Paper records are not always available at the time they are needed. Paper records are passive: they will not warn the physician if he overlooks some results. If the results of a lab request are unexpectedly not yet available, the paper record will not indicate that. Despite these drawbacks physicians are usually very positive about the use of the paper record, because they do not recognize their shortcomings.

Electronic patient records have some advantages over paper records. The legibility of electronic patient records is per definition good. Data can be presented according to different views (time-, source-, and problem-oriented), making the data easier to access. Electronic patient records, when interfaced with decision support systems, can provide reminders when a physician forgets something.

The use of computers in medical education is diverse. They can be used for managing the learning process, for distributing content, for assessing the student's knowledge or skills, and so on. In this case, they can be regarded as educational tools. As is clear from the above information systems are extensively used in the practice of health professionals. Students should be taught the benefits, but also the limitations of the use of information systems. Finally the information systems have to be developed. To be able to do so students need additional education in medical informatics.

BIBLIOGRAPHY

1. Institute of Medicine. To err is human: Building a safer health system. The National Academies Press; 2000.
2. Qayumi AK, et al. Comparison of computer-assisted instruction (CAI) versus traditional textbook methods for training in abdominal examination (Japanese experience). Med Ed 2004;38:1080–1088.
3. Hasman A, Haux R, Albert A. A systematic view on medical informatics. Comp Meth Prog Biomed 1996;51: 131–139.
4. Haux R, Grant A, Hasman A, Hovenga E, Knaup P. Recommendations of the International Medical Informatics Association (IMIA) on education in health and medical informatics. Methods Inf Med 2001;40:78–82.
5. http://www.hon.ch (last visited 21 December 2004).
6. van Merode GG, Groothuis S, Hasman A. Entreprise resource planning for hospitals. Int J Med Inform 2004;73: 493–501.
7. Fries JF. Alternatives in medical record formats. Med Care 1974;12:871–881.

MEDICAL ENGINEERING SOCIETIES AND ORGANIZATIONS

ARTHUR T JOHNSON
University of Maryland
College Park, Maryland

PATRICIA I HORNER
Landover, Maryland

INTRODUCTION

Modern technology has transformed the practice of medicine. We can now see where we could not before, conduct surgery with minimal trauma, intervene at the genetic level, replace whole natural organs with functional artificial ones, make rapid diagnoses, and peer into the workings of the brain. More patients are surviving, and those who do are living better. Much of the credit for these advances goes to the engineers, physicians, and physiologists who together decided what needed to be done, the science required to support it, and how it could be made practical. Medical engineers are now very much involved in the process of developing medical advances. They bring to medicine the abilities of conceptualization, computation, and commercialization. They use varied tools such as biophysics, applied mathematics, physiological modeling, bioinstrumentation and control, imaging, and biomechanics to accomplish their advances.

The result is that there are nearly as many subspecialties of medical engineering as there are medical specialties. Tissue engineers, for instance, grow bioartificial tissues and organs as replacements; metabolic engineers find means to adjust cellular metabolic pathways to produce greater quantities of biochemicals and hormones; and rehabilitation engineers design new prostheses or modify existing units to reestablish adequate function in patients who have lost ability usually as the result of trauma. There are medical engineers working with biosensors, bioprocess optimization, multiple imaging modes, pancreatic function, vascular replacement, and drug delivery. Biomaterials engineers have produced materials that can function in different regional corporal environments. Indeed, there is no part of the human body that has not been studied by medical engineers to improve or replace lost function.

As the body of medical knowledge has increased overall and has been repeatedly split more and more finely into specialties, there has been a concomitant proliferation of organizations to communicate, share, and advocate action related to their particular specialties. Some of these would be recognized as chiefly engineering organizations with application interests in medicine; some are medical societies with significant engineering contributions. There is almost no significant human disease, physiological system, organ, or function without a group or organization representing associated interests. There is even a group interested in developing synthetic biological forms that, although it is too premature to link with medicine, may someday have a profound effect on medicine. All of these groups can be found by searching the Internet, and any attempt to enumerate them here would be outdated very quickly.

DEFINITIONS

Progress in biological science and engineering has not been made with a clear distinction between medical and nonmedical applications. Advances in human medicine often find applications as well in veterinary medicine. Genetic coding techniques have been applied equally to humans and fruit flies. Prospective biomaterials are modeled on computer without regard for the ultimate specific application, and they are tested in animals, plants, or fungi before approval for human use. Progress toward better nutrition through science, and toward purer environments through improved pollutant detection monitoring, have resulted in better human health for most humans living in the developed world. Biology is biology, whether applied to human health care or not, so a convergence of basic knowledge and methods between medical and biological engineers is expected to continue.

Several relevant definitions attempt to distinguish among various fields where engineers have and will continue to contribute.

The U.S. National Institutes for Health (NIH) has the following definition of bioengineering:

> 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 levels; 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.

The U.S. National Science Foundation program in Biochemical Engineering and Biotechnology (BEB) describes its program in the following way:

> Advances the knowledge base of basic engineering and scientific principles of bioprocessing at both the molecular level (biomolecular engineering) and the manufacturing scale (bioprocess engineering). Many proposals supported by BEB programs are involved with the development of enabling technologies for production of a wide range of biotechnology products and services by making use of enzymes, mammalian, microbial, plant, and/or insect cells to produce useful biochemicals, pharmaceuticals, cells, cellular components, or cell composites (tissues).

The Whitaker Foundation definition of biomedical engineering is as follows:

> Biomedical engineering is a discipline that advances knowledge in engineering, biology, and medicine, and improves human health through cross-disciplinary activities that integrate the engineering sciences with the biomedical sciences and clinical practice. It includes: 1) The acquisition of new knowledge and understanding of living systems through the innovative and substantive application of experimental and analytical techniques based on the engineering sciences, and 2) The development of new devices, algorithms, processes, and systems that advances biology and medicine and improve medical practice and health care delivery.

And, finally, the Institute of Biological Engineering (IBE) defines biological engineering as follows:

> Biological engineering is the biology-based engineering discipline that integrates life sciences with engineering in the advancement and application of fundamental concepts of biological systems from molecular to ecosystem levels. The emerging discipline of biological engineering lies at the interfaces of biological sciences, engineering sciences, mathematics and computational sciences. It applies biological systems to enhance the quality and diversity of life.

HISTORICAL DEVELOPMENTS

In 1948, in New York City, a group of engineers from the Instrument Society of America (ISA) and the American Institute of Electrical Engineers (AIEE), with professional interests in the areas of X-ray and radiation apparatus used in medicine, held the First Annual Conference on Medical Electronics. Soon thereafter the Institute of Radio Engineers (IRE), joined with the ISA and AIEE, and the series of annual meetings continued. Subsequent years witnessed a remarkable growth of interest in biomedical engineering and participation by other technical associations. By 1968 the original core group evolved into the Joint Committee on Engineering in Medicine and Biology (JCEMB), with five adherent national society members: the Instrument Society of America (ISA), the Institute of Electrical and Electronics Engineers, Inc. (IEEE), the American Society of Mechanical Engineers (ASME), the American Institute of Chemical Engineers (AIChE), and the Association for the Advancement of Medical Instrumentation (AAMI), who jointly conducted the Annual Conference on Engineering in Medicine and Biology (ACEMB).

Professional groups responded vigorously to the demands of the times. Attendance at the annual conference by natural scientists and medical practitioners grew to approximately 40% of the total; medical associations requested formal participation with their technical counterparts on the JCEMB. New interdisciplinary organizations were formed. New intrasociety and intersociety groups, committees, and councils became active; meetings filled the calendar; and publications overflowed the shelves.

In 1968, a document was prepared that read as follows:
WHEREAS:

1. Common interdisciplinary purposes cannot be well served by individual groups working independently from each other;
2. Certain associations have developed in attempts to meet the need;
3. Conferences and publications have proliferated in attempts to meet the needs;
4. At present, no mutually satisfactory mechanism exists for the coordination of the relevant groups and functions;
5. There does exist an annual meeting and proceedings publication sponsored by a limited number of societies through the Joint Committee on Engineering in Medicine and Biology (JCEMB);
6. The JCEMB is formally structured with a constitution, plural societal representati9on, and an established pattern of operation. This structure and pattern of operation, however, are not deemed adequate to fulfill present and future needs. To the best of our knowledge, there exists no other single organization that seems capable of fulfilling these needs.

THEREFORE, it is appropriate that a new organization be established.

On July 21, 1969, at the 22nd ACEMB in Chicago, Illinois, representatives of 14 national engineering, scientific, and medical associations founded the Alliance for Engineering in Medicine and Biology (AEMB). It was incorporated on December 24, 1969, in Washington, D.C. Lester Goodman, Ph.D., served as Founder President in

1970–197l; Arthur C. Beall, MD(1972); Alan R. Kahn, MD(1973); Harry S. Lipscomb, MD(1974); Anthony Sances, Jr., Ph.D. (1975); Charles Weller, MD(1976–1977); Edward J. Hinman, MD MPH(1978–1979); Paul W. Mayer, MD(1980–1982); Francis M. Long, Ph.D., (1983–1984); Arthur T. Johnson, PE, Ph.D. (1985–1988); and Alfred R. Potvin, PE Ph.D. served as the final President in 1989–1990.

The Alliance operations were determined by an Administrative Council composed of delegates from each of its affiliates. Later the Alliance was to consist of more than 20 such organizations:

Aerospace Medical Association (ASMA)
American Academy of Orthopaedic Surgeons (AAOS)
American Association of Physicists in Medicine (AAPM)
American College of Chest Physicians (ACCP)
American College of Physicians (ACP)
American College of Radiology (ACR)
American College of Surgeons (ACP)
American Institute of Aeronautics and Astronautics (AIAA)
American Institute of Biological Sciences (AIBS)
American Institute of Chemical Engineers (AIChE)
American Institute of Ultrasound in Medicine (AIUM)
American Society for Artificial Internal Organs (ASAIO)
American Society for Engineering Education (ASEE)
American Society for Hospital Engineers of the American Hospital Association (ASHE)
American Society for Testing and Materials (ASTM)
American Society of Agricultural Engineers (ASAE)
American Society of Civil Engineers (ASCE)
American Society of Heating, Refrigerating and Air Conditioning Engineers (ASHRAE)
American Society of Internal Medicine (ASIM)
American Society of Mechanical Engineers (ASME)
Association for the Advancement of Medical Instrumentation (AAMI)
Biomedical Engineering Society (BMES)
Institute of Electrical and Electronics Engineers (IEEE)
Instrument Society of America (ISA)
National Association of Bioengineers (NAB)
Neuroelectric Society (NES)
RESNA—Rehabilitation Engineering & Assistive Technology Society of North America
Society for Advanced Medical Systems, now American Medical Informatics Association (AMIA)
Society for Experimental Stress Analysis (SESA)
SPIE—International Society for Optical Engineering
Alpha Eta Mu Beta—National Biomedical Engineering Student Honor Society, established under the auspices of A EMB.

The Alliance headquarters office opened on November 1, 1973. John H. Busser served as the first Executive Director. Patricia I. Horner served as Assistant Director, as Administrative Director, and succeeded Busser as the Executive Director. Among its goals, is the following excerpted in part from its constitution, bylaws, and recorded minutes:

> to promote cooperation among associations that have an active interest in the interaction of Engineering and the physical sciences with medicine and the biological sciences in enhancement of biomedical knowledge and health care.
>
> to establish an environment and mechanisms whereby people from relevant various disciplines can be motivated and stimulated to work together
>
> to respond to the needs of its member societies, as expressed by their delegates, rather than to seek authoritative preeminence in its domain of interest. . .
>
> to support and enhance the professional activities of its membership. . .

The 23rd ACEMB in Washington, D.C., in 1970, was the first held under the aegis of the Alliance. From 1979 to 1984, the IEEE Engineering in Medicine and Biology Society (EMBS) held their conferences immediately preceding the ACEMB. The Society for Advanced Medical Systems, later to become AMIA, and the Biomedical Engineering Society also held their meetings for several years in conjunction with the ACEMB.

The accomplishments of the Alliance far outstripped the expectations of its founders. The Alliance more than fulfilled responsibilities for the annual conference inherited from the predecessor JCEMB, but the Alliance made important contributions through a variety of studies and publications ranging from a 5-year ultrasound research and development agendum to a guideline for technology procurement in health-care institutions:

- First International Biomedical Engineering Workshop Series held in Dubrovnik, Yugoslavia, under the sponsorship of the National Science Foundation. This project was in cooperation with AIBS and the International Institute of Biomedical Engineering in Paris. Five workshops were held, and planning handbooks were completed.
- Assessment of selected medical instrumentation; Tasks 1–4, ultrasonic diagnostic imaging; Task 5, radiologic and radionuclide imaging technology.
- Summary guidelines and courses on technology procurement; practices and procedures for improving productivity in research and health-care institutions.
- Information exchange and problem assessments in medical ultrasound, including preparation and distribution of a directory of federal activities, conducted instrumentation conferences, delineated training needs, assessed technology transfer potential, and prepared guidelines for the establishment of clinical ultrasound facilities.

- Joint U.S.–Egypt international technology transfer project in medical diagnostic ultrasound, including international workshops and the design and support of a focus laboratory for ultrasonic diagnosis at Cairo University Medical School.
- Short courses for continuing education at the annual conference on engineering in medicine and biology.
- International directory of biomedical engineers.

Before long, the proliferation of medical engineers, and competing interests among societies, led to a fragmentation of the field. It became clear that the Alliance no longer represented positions of the entire field. No organized group could speak for the entire profession, and the spirit of unity that had led to the development of AEMB no longer existed. It was time for a new beginning.

AMERICAN INSTITUTE FOR MEDICAL AND BIOLOGICAL ENGINEERING

In 1988, the National Science Foundation funded a grant to develop an infrastructure for bioengineering in the United States. The AEMB, jointly with the U.S. National Committee on Biomechanics (USNCB), was to develop a unifying organization for bioengineering in the United States. The co-principal investigators were Robert M. Nerem and Arthur T. Johnson, and Patricia Horner served as Project Director. The AEMB/USNCB Steering Committee consisted of Robert M. Nerem, Arthur T. Johnson, Michael J. Ackerman, Gilbert B. Devey, Clifford E. Brubaker, Morton H. Friedman, Dov Jaron, Winfred M. Phillips, Alfred R. Potvin, Jerome S. Schultz, and Savio L-Y Woo. The Steering Committee met in January and March 1989, and the first workshop was held in August 1989. Two more Steering Committee meetings were held in December 1989 and March 1990, and the second workshop was held in July 1990. The outcome of these two workshops was to establish the American Institute for Medical and Biological Engineering (AIMBE). All AEMB members voted to cease operation of the Alliance for Engineering in Medicine and Biology in 1990 and to transfer the AEMB assets and 501(c)3 status to AIMBE in 1991.

AIMBE opened an office in Washington, D.C., in 1995 with Kevin O'Connor as Executive Director. He was succeeded by Arthur T. Johnson in 2004 and Patricia Ford-Roegner in 2005. AIMBE Presidents have been as follows: Robert Nerem (1992–1993), Pierre Galletti (1994), Jerome Schultz (1995), Winfred Phillips (1996), Larry McIntire (1997), William Hendee (1998), John Linehan (1999), Shu Chien (2000), Peer Portner (2001), Buddy Ratner (2002), Arthur Coury (2003), Don Giddens (2004), Thomas Harris (2005), and Herbert Voigt (2006).

Representing over 75,000 bioengineers, the AIMBE seeks to serve and coordinate a broad constituency of medical and biological scientists and practitioners, scientific and engineering societies, academic departments, and industries. Practical engagement of medical and biological engineers within the AIMBE ranges from the fields of clinical medicine to food, agriculture, and environmental bioremediation.

AIMBE's mission is to

- Promote awareness of the field and its contributions to society in terms of new technologies that improve medical care and produce more and higher quality food for people throughout the world.
- Work with lawmakers, government agencies, and other professional groups to promote public policies that further advancements in the field.
- Strive to improve intersociety relations and cooperation within the field.
- Promote the national interest in science, engineering, and education.
- Recognize individual and group achievements and contributions to medical and biological engineering.

AIMBE is composed of four sections:

- The College of Fellows—1000 Persons who are the outstanding bioengineers in academic, industry, and government. These leaders in the field have distinguished themselves through their contributions in research, industrial practice, and/or education. Most Fellows come from the United States, but there are international Fellows.
- The Academic Council—Universities with educational programs in bioengineering at the graduate or undergraduate level. Currently there are approximately 85 member institutions. Representative to the Council generally are chairs of their departments. Many also are members of the College of Fellows. The Council considers issues ranging from curricular standards and accreditation to employment of graduates and funding for graduate study.
- The Council of Societies—The AIMBE's mechanism coordinating interaction among 19 scientific organizations in medical and biological engineering. The purposes of the Council are to provide a collaborative forum for the establishment of society member positions on issues affecting the field of medical and biological engineering, to foster intersociety dialoge and cooperation that provides a cohesive public representation for medical and biological engineering, and to provide a way to coordinate activities of member societies with the activities of academia, government, the health-care sector, industry, and the public and private biomedical communities.
- The Industry Council—A forum for dialog among industry, academia, and government to identify and act on common interests that will advance the field of medical and biological engineering and contribute to public health and welfare. Industrial organizations may be members of the Industry Council if they have substantial and continuing professional

interest in the field of medical and biological engineering.

Current members of the Council of Societies are as follows:

American Association of Physicists in Medicine
American College of Clinical Engineering
American Institute of Chemical Engineers; Food, Pharmaceutical and Bioengineering Division
American Medical Informatics Association
American Society of Agricultural and Biological Engineers
American Society for Artificial Internal Organs
American Society for Biomechanics
American Society of Mechanical Engineers, Bioengineering Division
Biomedical Engineering Society
Controlled Release Society
IEEE Engineering in Medicine and Biology Society
Institute of Biological Engineering
International Society for Magnetic Resonance in Medicine
Orthopaedic Research Society
Rehabilitation Engineering and Assistive Technology Society of North America
Society for Biomaterials
SPIE: The International Society for Optical Engineering
Surfaces in Biomaterials Foundation

Current members of the Industry Council are as follows:

Biomet, Inc.
Boston Scientific Corporation
Genzyme Corporation
Medtronic, Inc.
Pequot Ventures
Smith + Nephew
Vyteris, Inc.
Wright Medical Technology, Inc.
Zimmer, Inc.

The AIMBE Board of Directors oversees the work of the College of Fellows and the three councils. The Board consists of a President who is assisted by two Past Presidents, the President-Elect, four Vice-Presidents at Large, a Secretary-Treasurer, and the Chair of the College of Fellows—all of whom are elected by the Fellows. The Board also includes chairs of the other councils and chairs of all standing committees. AIMBE's day-to-day operations are supervised by the Executive Director in the Washington headquarters.

AIMBE's Annual Event each winter in Washington, D.C., provides a forum on the organization's activities and is a showcase for key developments in medical and biological engineering. The annual event includes a 1-day scientific symposium sponsored by the College of Fellows, a ceremony to induct the newly elected Fellows, and a 1-day series of business meetings focused on public policy and other issues of interest to AIMBE's constituents. For additional information about AIMBE's mission, memberships, and accomplishments, visit http://www.aimbe.org.

The AIMBE has focused on public policy issues associated with medical and biological engineering. The AIMBE enjoys high credibility and respect based on the stature of its Fellows, support from constituent societies, and its intention to be a forum for the best interests of the entire field. The AIMBE has taken positions on several important issues and advocated that they be adopted by various agencies and by Congress. A few of the AIMBE'S public policy initiatives that have met with success are as follows:

- National Institute of Biomedical Imaging and Bioengineering (NIBIB)—Created in 2000 with the help of AIMBE advocacy, the NIBIB has received strong support from the AIMBE and other institutions that value the role of technology in medicine, particularly the Academy of Radiological Research. The NIBIB has experienced rapid growth and development in all areas, including scientific programs, science administration, and operational infrastructure. The prognosis for the near future is continued growth and development especially in bioengineering, imaging, and interdisciplinary biomedical research and training programs.
- FDA Modernization Act (FDAMA)—Enacted in 1997, this legislation amended the Federal Food, Drug, and Cosmetic Act relation to the regulation of food, drugs, devices, and biological products. FDAMA enhanced the FDA's mission in ways that recognized the Agency would be operating in a twenty-first century characterized by increasing technological, trade, and public health complexities.
- Biomaterials Access Assurance Act—The 1998 legislation provides relief for materials suppliers to manufacturers of implanted medical devices by allowing those suppliers to be dismissed from lawsuits in which they are named if they meet the statutory definition of a "biomaterials supplier."
- National Institutes of Health Bioengineering Consortium (BECON)—This is the focus of bioengineering activities at the NIH. The Consortium consists of senior-level representatives from all NIH institutes, centers, and divisions plus representatives of other Federal agencies concerned with biomedical research and development. The BECON is administered by NIBIB.

The AIMBE Hall of Fame was established in 2005 to recognize and celebrate the most important medical and biological engineering achievements contributing to the quality of life. The Hall of Fame provides tangible evidence of the contributions of medical and biological engineering during the following decades:

1. *1950s and earlier*
   - Artificial kidney
   - X ray
   - Cardiac pacemaker
   - Cardiopulmonary bypass
   - Antibiotic production technology
   - Defibrillator
2. *1960s*
   - Heart valve replacement
   - Intraocular lens
   - Ultrasound
   - Vascular grafts
   - Blood analysis and processing
3. *1970s*
   - Computer-assisted tomography (CT)
   - Artificial hip and knee replacement
   - Balloon catheter
   - Endoscopy
   - Biological plant/food engineering
4. *1980s*
   - Magnetic resonance imaging (MRI)
   - Laser surgery
   - Vascular stents
   - Recombinant therapeutics
5. *1990s*
   - Genomic sequencing and micro-arrays
   - Positron emission tomography
   - Image-guided surgery

The AIMBE has now turned its attention to Barriers to Further Innovation. It is providing forums and platforms for identification and discussion of obstacles standing in the way of advances in medical and biological engineering. Barriers could be procedures, policies, attitudes, or information and education, anything that can yield when AIMBE constituents apply pressure at appropriate levels.

OTHER SOCIETIES

These are other general biomedical engineering societies that operate within the United States Among these, the Biomedical Engineering Society (BMES), Engineering in Medicine and Biology Society (EMBS), and the Institute for Biological Engineering (IBE) are probably the most inclusive. Others direct their attentions to specific parts of the discipline. There are trade organizations that have an industry perspective (such as AdvaMed for the medical device industry and the Biotechnology Industry Organization (BID) for the biotech industry), and there are peripheral organizations that deal with public health, the environment, and biotechnology. Many of these organizations publish excellent journals, newsletters, and information sheets. Those from trade organizations are often distributed free of charge, but they do not include peer-reviewed articles. Information about these can be found on the Internet.

Internationally, an organizational hierarchy exists. National and transnational organizations can belong to the International Federation for Medical and Biological Engineering (IFMBE), and that confers membership privileges to all AIMBE members and constituent society members. The IFMBE and the International Organization for Medical Physics (IOMP) together jointly sponsor a World Congress on Medical Physics and Biomedical Engineering every 3 years. The IOMP and IFMBE are members of the International Union for Physical and Engineering Sciences in Medicine (IUPESM), and the IUPESM, in turn, is a member of the International Council for Science (ICSU). ICSU members are national and international scientific unions and have a very broad and global outreach.

THE FUTURE

At least for the foreseeable future, new groups will be formed representing medical engineering specialties. Whether these groups organize formally and persist will depend on the continuing importance of their areas of focus. The organizations with a more general foci will continue to function and may spawn splinter groups. Given the political importance of concerted effort, organizations such as the AIMBE will continue to be active in promoting policy. Competitive pressures among different organizations, especially when expectations of continuing growth cannot be sustained, will always be a threat to the current order. Given that the cycle of competition and disorder leading to a realization that some ordered structure is preferable has been repeated at least once, there will continue to be some undercurrent of turmoil within the community of medical engineering organizations.

U.S. PROFESSIONAL SOCIETIES AND ORGANIZATIONS

Biomedical Engineering Associations and Societies

Advamed. 1200 G Street NW, Suite 400, Washington, D.C. 20005. 202-783-8700, http://www.advamed.org. Stephen J. Ubl, President. 1300 Members.

Represents manufacturers of medical devices, diagnostic products, and medical information systems. AdvaMed's members manufacture nearly 90% of the $80 billion of health-care technology purchased annually in the United States. Provides advocacy, information, education, and solutions necessary for success in a world of increasingly complex medical regulations.

Alpha Eta Mu Beta. 8401 Corporate Drive, Suite 140, Landover, MD 20785. 301-459-1999, http://www.ahmb.org.

Herbert F. Voigt, National President; Patricia I. Horner, Executive Director. 20 chapters.

Alpha Eta Mu Beta, the National Biomedical Engineering Honor Society, was founded by Daniel Reneau at Louisiana Tech University in 1979. This organization was sponsored by the AEMB. The AEMB was established to mark in an outstanding manner those biomedical engineering students who manifested a deep interest and marked ability in their chosen life work to promote an understanding of their profession and to develop its members professionally.

American Institute for Medical and Biological Engineering. 1901 Pennsylvania Ave NW, Suite 401, Washington, D.C. 20006. 202-496-9660, http://www.aimbe.org. Patricia Ford Roegner, Executive Director. 1000 Fellows; 18 Scientific Organizations; 85 Universities.

Founded in 1991 to establish an identity for the field of medical and biological engineering, which is the bridge between the principles of engineering science and practice and the problems and issues of biological and medical science and practice. The AIMBE comprises four sections. The College of Fellows with over 1000 persons who are the outstanding bioengineers in academia, industry, and government. The Academic Council is 85 universities with educational programs in bioengineering at the graduate or undergraduate level. The Council of Societies is 18 scientific organizations in medical and biological engineering. The Industry Council is a forum for dialog among industry, academia, and government. Principal activities include participation in formulation of public policy, dissemination of information, and education. Affiliated with the International Federation for Medical and Biological Engineering. Annual event each winter in Washington, D.C.

American Association of Engineering Societies. 1828 L Street NW, Suite 906, Washington, D.C. 20036. 202-296-2237, http://www.aaes.org. Thomas J. Price, Executive Director. 26 Engineering Societies.

Founded in 1979 in New York City. Member societies represent the mainstream of U.S. engineering with more than one million engineers in industry, government, and academia. The AAES has four primary programs: communications, engineering workforce commission, international, and public policy. Governance consists of two representatives from each of 26 member societies. Convenes diversity summits, publishes engineering and technology degrees, and holds annual awards ceremony.

American Academy of Environmental Engineers. 130 Holiday Court, Suite 100, Annapolis, MD 21401. 410-266-3311, http://www.aaee.net. David A. Asselin, Executive Director.

The American Sanitary Engineering Intersociety Board incorporated in 1955 became the American Academy of Environmental Engineers in 1966; and in 1973, it merged with the Engineering Intersociety Board. Principal purposes are improving the practice, elevating the standards, and advancing public recognition of environmental engineering through a program of specialty certification of qualified engineers.

American Academy of Orthopaedic Surgeons. 600 North River Road, Rosemont, IL 60018. 847-823-7186, http://www.aaos.org. Karen L. Hackett, Chief Executive Officer. 24,000 Members.

Founded in Chicago in 1933. Provides education and practice management services for orthopedic surgeons and allied health professionals. Maintains a Washington, D.C., office. Annual spring meeting.

American Academy of Orthotists and Prosthetists. 526 King Street, Suite 201, Alexandria, VA 22314. 703-836-0788, http://www.oandp.org. Peter D. Rosenstein, Executive Director. 3000 Members.

Founded in 1970 to further the scientific and educational attainments of professional practitioners in the disciplines of orthotics and prosthetics. Members have been certified by the American Board for Certification in Orthotics and Prosthetics. Annual spring meeting.

American Association of Physicists in Medicine. One Physics Ellipse, College Park, MD 20740. 301-209-3350, http://www.aapm.org. Angela R. Keyser, Executive Director. 4700 Members.

Founded in Chicago in 1958 and incorporated in Washington in 1965. Promotes the application of physics to medicine and biology. Member society of the American Institute of Physics. Annual summer meeting.

American Chemical Society. 1155 Sixteenth Street NW, Washington, D.C. 20036. 800-227-5558, http://www.chemistry.org. Madeleine Jacobs, Executive Director, and CEO. 159,000 Members.

Founded in 1877 in New York City. Granted a national charter by Congress in 1937. Encourages the advancement of chemistry. Semiannual spring and fall meetings.

American College of Nuclear Physicians. 1850 Samuel Morse Drive, Reston, VA 20190. 703-326-1190, http://www.acnponline.org. Virginia M. Pappas, Executive Director. 500 Members.

Established in 1974, the organization provides access to activities that encompass the business and economics of nuclear medicine as they impact nuclear medicine physicians. Semiannual meetings fall and winter.

American College of Physicians. 190 N. Independence Mall West, Philadelphia, PA 19106. 215-351-2600, http://www.acponline.org. John Tooker, Executive Vice President. 119,000 Members.

Founded in New York City in 1915. Merged with the Congress of Internal Medicine in 1923 and merged in 1998 with the American Society of Internal Medicine. Patterned after England's Royal College of Physicians. Members are physicians in general internal medicine and related subspecialties. Maintains a Washington, D.C., office. Annual spring meeting.

American College of Radiology. 1891 Preston White Drive, Reston, VA 20191. 703-648-8900, http://www.acr.org. Harvey L. Neiman, Executive Director. 30,000 Members.

Founded in 1923 in San Francisco and incorporated in California in 1924. Purpose is to improve the health of patients and society by maximizing the value of radiology and radiologists by advancing the science, improving patient service, and continuing education. Annual fall meeting.

American College of Surgeons. 633 N. St. Clair Street, Chicago, IL 60611. 312-202-5000, http://www.facs.org. Thomas R. Russell, Executive Director. 64,000 Fellows, 5000 Associate Fellows.

Founded in 1913 and incorporated in Illinois. U.S. member of the International Federation of Surgical Colleges. Members are Fellows who must meet high standards established by the College. Purpose is to improve the quality of care for the surgical patient by setting high standards for surgical education and practice. Annual fall clinical congress.

American Congress of Rehabilitation Medicine. 6801 Lake Plaza Drive, Suite B-205, Indianapolis, IN 46220. 317-915-2250, http://www.acrm.org. Richard D. Morgan, Executive Director. 1700 Members, 15 Companies.

Founded in 1923 as the American College of Radiology and Physiotherapy. Name changed in 1926 to American College of Physical Therapy and in 1930 to American Congress of Physical Therapy. Changed again in 1945 to American Congress of Physical Therapy and in 1953 became American Congress of Physical Medicine and Rehabilitation. Adopted its current name in 1967. Provides education for professionals in medical rehabilitation. Fall annual meeting.

American Institute of Biological Sciences. 1444 I Street NW, Suite 200, Washington, D.C. 20005. 202-628-1500, http://www.aibs.org. Richard O'Grady, Executive Director. 80 Societies, 6000 Members.

Founded in 1947 as part of the National Academy of Sciences. Incorporated as an independent nonprofit since 1954. Absorbed America Society of Professional Biologists in 1969. Represents more than 80 professional societies with combined membership exceeding 240,000 scientists and educators. Also more than 6000 individual members. Purpose is to better serve science and society. Annual meeting in August.

American Institute of Chemical Engineers. 3 Park Avenue, New York, NY 10016. 212-591-8100, http://www.aiche.org. John Sofranko, Executive Director. 40,000 Members.

Organized in 1908 and incorporated in 1910. Member of Accreditation Board for Engineering and Technology, American National Standards Institute, American Association of Engineering Societies, and other related organizations. Purpose is to advance the chemical engineering profession. Annual meeting in November.

American Institute of Physics. One Physics Ellipse, College Park, MD 20740. 301-209-3131, http://www.aip.org. Marc H. Brodsky, Executive Director, and Chief Executive Officer. 10 Societies and 24 Affiliates.

Chartered in 1931 to promote the advancement of physics and its application to human welfare. Federation of 10 Member Societies representing spectrum of physical sciences.

American Institute of Ultrasound in Medicine. 14750 Sweitzer Lane, Suite 100, Laurel, MD 20707. 301-498-4100, http://www.aium.org. Carmine Valente, Chief Executive Officer.

Began in 1951 at a meeting of 24 physicians attending the American Congress of Physical Medicine and Rehabilitation. Membership includes biologists, physicians, and engineers concerned with the use of ultrasound for diagnostic purposes. Provides continuing education, CME tests, and accreditation of ultrasound laboratories. Annual fall meeting.

American Medical Informatics Association. 4915 St. Elmo Avenue, Suite 401, Bethesda, MD 20814. 301-657-1291, http://www.amia.org. Don Detmer, President, and CEO. 3000 Members.

Founded in 1990 through a merger of three existing health informatics associations. Members represent all basic, applied, and clinical interests in health-care information technology. Promotes the use of computers and information systems in health care with emphasis on direct patient care. Semiannual meetings: spring congress in the West and fall annual symposium in the East.

American Society for Artificial Internal Organs. P.O. Box C, Boca Raton, FL 33429. 561-391-8589, http://www.asaio.net. 1400 Members.

Established in 1955 in Atlantic City, NJ. Annual June conference.

American Society for Engineering Education. 1818 N Street NW, Suite 600, Washington, D.C. 20036. 202-331-3500, http://www.asee.org. Frank L. Huband, Executive Director. 12,000 Members, 400 Colleges, 50 Corporations.

Founded in 1893 as the Society for Promotion of Engineering Education. Incorporated in 1943 and merged in 1946 with Engineering College Research Association. Members include deans, department heads, faculty members, students, and government and industry representatives from all disciplines of engineering and engineering technology. Member of the American Association of Engineering Societies, Accreditation Board for Engineering and Technology, American Institute for Medical and Biological Engineering, and American Council on Education. Participating society of World Federation of Engineering Associations. Purpose is to further education in engineering and engineering technology. Annual June meeting.

American Society for Healthcare Engineering of the American Hospital Association. One North Franklin, 28th Floor, Chicago, IL 60606. 312-422-3800, http://www.ashe.org. Albert J. Sunseri, Executive Director.

Affiliate of the American Hospital Association. Annual June meeting.

American Society for Laser Medicine and Surgery. 2404 Stewart Avenue, Wausau, WI 54401. 715-845-9283, http://www.aslms.org. Dianne Dalsky, Executive Director. 3000 Members.

Founded in 1980 to promote excellence in patient care by advancing laser applications and related technologies. Annual spring meeting.

American Society of Agricultural and Biological Engineers. 2950 Niles Road, St. Joseph, MI 49085. 269-429-0300, http://www.asabe.org. Melissa Moore, Executive Vice President. 9000 Members.

Founded in 1907 as the American Society of Agricultural Engineers and changed its name in 2005. Dedicated to the advancement of engineering applicable to agricultural, food, and biological systems. Annual meeting in July.

American Society of Civil Engineers. 1801 Alexander Bell Drive, Reston, VA 20191. 703-295-6000, http://www.asce.org. Patrick J. Natale, Executive Director. 137,500 Members.

Founded in 1852 as the American Society of Civil Engineers and Architects. Dormant from 1855 to 1867, but it revived in 1868 and incorporated in 1877 as the American Society of Civil Engineers. Over 400 local affiliates, 4 Younger Member Councils, 230 Student Chapters, 36 Student Clubs, and 6 International Student Groups. Semiannual spring and fall meetings.

American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc. 1791 Tullie Circle NE, Atlanta, GA 30329. 404-636-8400, http://www.ashrae.org.

Incorporated in 1895 as the American Society of Heating and Ventilating Engineers, known after 1954 as American Society of Heating and Air-Conditioning Engineers. Merged in 1959 with American Society of Refrigerating Engineers to form American Society of Heating, Refrigerating and Air-Conditioning Engineers. Annual summer meeting.

American Society of Mechanical Engineers. Three Park Avenue, New York, NY 10016. 212-591-7722, http://www.asme.org. Virgil R. Carter, Executive Director. 120,000 Members.

Founded in 1880 and incorporated in 1881. Focuses on technical, educational, and research issues of engineering and technology. Sets industrial and manufacturing codes and standards that enhance public safety. Conducts one of the world's largest technical publishing operations. Semiannual summer and winter meetings.

American Society of Neuroradiology. 2210 Midwest Road, Suite 207, Oak Brook, IL 60523. 630-574-0220, http://www.asnr.org. James B. Gantenberg, Executive Director/CEO. 2700 Members.

Founded in 1962. Supports standards for training in the practice of neuroradiology. Annual spring meeting.

American Society of Safety Engineers. 1800 E. Oakton Street, Des Plaines, IL 60018. 847-699-2929, http://www.asse.org. Fred Fortman, Executive Director. 30,000 Members.

Founded in 1911 as the United Association of Casualty Inspectors and merged with the National Safety Council in 1924, becoming its engineering section. Became independent again in 1947 as the American Society of Safety Engineers and incorporated in 1962. There are 13 practice specialties, 150 chapters, 56 sections, and 64 student sections. Annual spring meeting.

Association for Computing Machinery. 1515 Broadway, New York, NY 10036. 212-626-0500, http://www.acm.org. John R. White, Executive Director. 80,000 Members.

Founded in 1947 at Columbia University as Eastern Association for Computing Machinery and incorporated in Delaware in 1954. Affiliated with American Association for Advancement of Science, American Federation of Information Processing Societies, Conference Board of Mathematical Sciences, National Academy of Sciences-National Research Council, and American National Standards Institute. Advancing the skills of information technology professionals and students. Annual fall meeting.

Association for the Advancement of Medical Instrumentation. 1100 North Glebe Road, Suite 220, Arlington, VA 22201. 703-525-4890, http://www.aami.org. Michael J. Miller, President. 6000 Members.

Founded in 1967, the AAMI is an alliance of over 6000 members united by the common goal of increasing the understanding and beneficial use of medical instrumentation and technology. Annual spring meeting.

Association of Biomedical Communications Directors. State University of New York at Stony Brook, Media Services L3044 Health Sciences Center, Stony Brook, NY 11794. 631-444-3228. Kathleen Gebhart, Association Secretary. 100 Members.

Formed in 1974 as a forum for sharing information; adopted a Constitution and Bylaws in 1979, and incorporated in April 1979 in North Carolina. Members are directors of biomedical communication in academic health science settings. Annual spring meeting.

Association of Environmental Engineering and Science Professors. 2303 Naples Court, Champaign, IL 61822. 217-398-6969, http://www.aeesp.org. Joanne Fetzner, Secretary. 700 Members.

Formerly, in 1972, the American Association of Professors in Sanitary Engineering. Professors in academic programs throughout the world who provide education in the sciences and technologies of environmental protection. Biennial conference in July.

Biomedical Engineering Society. 8401 Corporate Drive, Suite 140, Landover, MD 20785. 301-459-1999, http://www.bmes.org. Patricia I. Horner, Executive Director. 3700 Members.

Founded in 1968 in response to a need to give equal representation to both biomedical and engineering interests. The purpose of the Society is to promote the increase of biomedical engineering knowledge and its use. Member of American Institute for Medical and Biological Engineering. Annual fall meeting.

Biophysical Society. 9650 Rockville Pike, Bethesda, MD 20814. 301-634-7114, http://www.biophysics.org. Ro Kampman, Executive Officer. 7000 Members.

Founded in 1957 in Columbus, OH, to encourage development and dissemination of knowledge in biophysics. Annual winter meeting.

Health Physics Society. 1313 Dolley Madison Boulevard, Suite 402, McLean, VA 22101. 703-790-1745, http://www.hps.org. Richard J. Burk, Jr, Executive Secretary. 7000 Members.

Founded in 1956 in the District of Columbia and reincorporated in Tennessee in 1969. Society specializes in occupational and environmental radiation safety. Affiliated with International Radiation Protection Association. Annual summer meeting.

Human Factors and Ergonomics Society. P.O. Box 1369, Santa Monica, CA 90406. 310-394-1811, http://www.hfes.org. Lynn Strother, Executive Director. 50 Active Chapters and 22 Technical Groups.

Founded in 1957, formerly known as the Human Factors Society. An interdisciplinary organization of professional people involved in the human factors field. Member of the International Ergonomics Association. Annual meeting in September-October.

Institute for Medical Technology Innovation. 1319 F Street NW, Suite 900, Washington, D.C. 20004. 202-783-0940, http://www.innovate.org. Martyn W.C. Howgill, Executive Director.

The concept was developed in 2003 by leaders in the medical device industry, and the Institute was incorporated and opened its offices in 2004. Purpose is to demonstrate the role, impact, and value of medical technology on health care, economy, and society, for the benefit of patients.

Institute of Electrical and Electronics Engineers. 3 Park Avenue, 17th Floor, New York, NY 10016. 212-419-7900, http://www.ieee.org. Daniel J. Senese, Executive Director. 365,000 Members.

The American Institute of Electrical Engineers was founded in 1884 and merged in 1963 with the Institute of Radio Engineers. Three Technical Councils, 300 local organizations, 1300 student branches at universities, and 39 IEEE Societies including the Engineering in Medicine and Biology Society with 8000 members and meets annually in the fall. Maintains Washington, D.C. office.

Institute of Environmental Sciences and Technology. 5005 Newport Drive Suite 506, Rolling Meadows, IL 60008. 847-255-1561, http://www.iest.org. Julie Kendrick, Executive Director.

Formed by a merger of the Institute of Environmental Engineers and the Society of Environmental Engineers in 1953. Annual spring meeting.

Instrument Society of America. 67 Alexander Drive, Research Triangle Park, NC 27709. 919-549-8411, http://www.isa.org. Rob Renner, Executive Director. 30,000 Members.

Founded in Pittsburgh in 1945. Charter member of American Automatic Control Council, affiliate of American Institute of Physics, member of American Federation of Information Processing Societies, member of American National Standards Institute, and U.S. representative to the International Measurement Confederation. Develops standards, certifies industry professionals, provides education and training, publishes books and technical articles, and hosts largest conference for automation professionals in the Western Hemisphere. Annual meeting in October.

International Biometric Society ENAR. 12100 Sunset Hills Road, Suite 130, Reston, VA 22090. 703-437-4377, http://www.enar.org. Kathy Hoskins, Executive Director. 6500 Members.

Founded in September 1947. Became the International Biometric Society in 1994. Annual March and June meetings.

International College of Surgeons. 1516 North Lake Shore Drive, Chicago, IL 60610. 312-642-3555, http://www.icsglobal.org. Max C. Downham, Executive Director. 10,000 Members.

Founded in Geneva, Switzerland, in 1935 and incorporated in the District of Columbia in 1940. Federation of general surgeons and surgical specialists. Annual spring meeting of U.S. section and biennial international meetings.

International Society for Magnetic Resonance in Medicine. 2118 Milvia Street, Suite 201, Berkeley, CA 94704. 510-841-1899, http://www.ismrm.org. Roberta A. Kravitz, Executive Director. 6,000 Members.

Formed as a merger of the Society for Magnetic Resonance Imaging and Society of Magnetic Resonance in Medicine in 1995. Promotes the application of magnetic resonance techniques to medicine and biology. Annual meetings in April/May.

Medical Device Manufacturers Association. 1919 Pennsylvania Avenue NW, Suite 660, Washington, D.C. 20006. 202-349-7171, http://www.medicaldevices.org. Mark B. Leahey, Executive Director. 140 Companies.

Created in 1992. Supersedes Smaller Manufacturers Medical Device Association. Represents manufacturers of medical devices. Annual May meeting.

Radiation Research Society. 810 East 10th Street, Lawrence, KS 666044. 800-627-0629, http://www.radres.org. Becky Noordsy, Executive Director. 2025 Members.

Founded in 1952 as a professional society of persons studying radiation and its effects. Annual spring-summer meetings.

Radiological Society of North America. 820 Jorie Boulevard, Oak Brook, IL 60523. 630-571-2670, http://

www.rsna.org. Dave Fellers, Executive Director. 37,577 Members.

Founded as Western Roentgen Society and assumed its current name in 1918. Members are interested in the application of radiology to medicine. Holds the largest medical meeting in the world annually in November with more than 60,000 in attendance.

RESNA—Rehabilitation Engineering & Assistive Technology Society of North America. 1700 N. Moore Street, Suite 1540, Arlington, VA 22209. 703-524-6686, http://www.resna.org. Larry Pencak, Executive Director. 1000 Members.

Founded in 1979 as the Rehabilitation Engineering Society of North America. In June 1995, the name was changed to the Rehabilitation Engineering and Assistive Technology Society of North American—RESNA. Twenty-one special interest groups and seven professional specialty groups. Annual meeting in June.

SPIE—International Society for Optical Engineering. P.O. Box 10, Bellingham, WA 98227. 360-676-3290, http://www.spie.org. Eugene G. Arthurs, Executive Director. 14,000 Members, 320 Companies.

Founded in 1956 in California as the Society of Photographic Instrumentation Engineers, it later became the Society of Photo-Optical Instrumentation Engineers and assumed its current name in 1981. Members are scientists, engineers, and companies interested in application of optical, electro-optical, fiber-optic, laser, and photographic instrumentation systems and technology. Semiannual meetings.

Society for Biological Engineering of the American Institute of Chemical Engineers. 3 Park Avenue, New York, NY 10016. 212-591-7616, http://www.bio.aiche.org.

Established by the AIChE for engineers and applied scientists integrating biology with engineering.

Society for Biomaterials. 15000 Commerce Parkway, Suite C, Mt. Laurel, NJ 08054. 856-439-0826, http://www.biomaterials.org. Victoria Elliott, Executive Director. 2100 Members.

Founded in 1974. Promotes biomaterials and their uses in medical and surgical devices. Annual spring meeting.

Society for Biomolecular Screening. 36 Tamarack Avenue, #348, Danbury, CT 06811. 203-743-1336, http://www.sbsonline.org. Christine Giordano, Executive Director. 1080 Members, 230 Companies.

Supports research in pharmaceutical biotechnology and the agricultural industry that use chemical screening procedures. Annual fall meeting.

Society for Modeling and Simulation International. P.O. Box 17900, San Diego, CA 92177. 858-277-3888, http://www.scs.org. Steve Branch, Executive Director.

Established in 1952 as the Simulation Council and incorporated in California in 1957 as the Simulation Councils. Became Society for Computer Simulation in 1973 and later changed its name to the current one. A founding member of the Information Processing Societies and National Computer Confederation Board, and affiliated with American Association for the Advancement of Science. Holds regional simulation multiconferences.

Society of Interventional Radiology. 3975 Fair Ridge Drive, Suite 400 North, Fairfax, VA 22033. 703-691-1805, http://www.sirweb.org. Peter B. Lauer, Executive Director.

Society of Nuclear Medicine. 1850 Samuel Morse Drive, Reston, VA 20190. 703-709-9000, http://www.interactive.snm.org. Virginia Pappas, Executive Director.

Society of Rheology. Suite 1N01, 2 Huntington Quadrangle, Melville, NY 11747. 516–2403, http://www.rheology.org. Janis Bennett, Executive Director.

Permanent address is at the American Institute of Physics and is one of five founding members of the AIP. Members are chemists, physicists, biologists, and others concerned with theory and precise measurement of flow of matter and response of materials to mechanical force. Annual meeting held in October or November.

Professional Societies and Organizations of Other Countries

Pick up from IFMBE Affiliates on www.ifmbe.org.

Further Reading

Biomedical Engineers, Brief 519, G.O.E. 02.02.01; D.O.T. (4th ed.) 019. Chronicle Guidance Publications, Moravia, NY 13118, 1994.

Biomedical Engineers. *Occupational Outlook Handbook, 2004-05 Edition*. Bureau of Labor Statistics, US. Department of Labor. http://www.bls.gov/oco/ocos262.htm.

Boykin D. Biomedical engineering takes center stage. Engineering Times 2004; 26 (9). National Society of Professional Engineers, 1420 King Street, Alexandria, VA 22314.

Collins CC. The retrospectroscope: Notes on the history of biomedical engineering in America. IEEE Eng Med Biol Mag Dec. 1988. IEEE Engineering in Medicine & Biology Society, 445 Hoes Lane, Piscataway, NJ 08854.

Enderle J, editor. Charting the milestones of biomedical engineering. *IEEE Eng Med Biol Mag*, May 2002. IEEE Engineering in Medicine and Biology Society, 445 Hoes Lane, Piscataway, NJ 08854.

Fagette PH je, Homer PI, editor. The Biomedical Engineering Society: An Historical Perspective. Landover, MD: The Biomedical Engineering Society; 2004.

Goodman L. The International Federation for Medical and Biological Engineering—20 years on. Med Biol Eng Comput Jan. 1978. International Federation for Medical and Biological Engineering.

Katona P. The Whitaker Foundation: The end will be just the beginning. IEEE Trans Med Imaging 2002;21(8). IEEE Engineering in Medicine & Biology Society, 445 Hoes Lane, Piscataway, NJ 08854.

Johnson AT. Executive Director's Report: What is ASMBE all about? ASMBE Newslett 2004:1.

Planning a Career in Biomedical Engineering (2004). Biomedical Engineering Society, 8401 Corporate Drive, Suite 140, Landover, MD 20785. http://www.bmes.org

Tompkins W. From the President. IEEE Eng Med Biol Mag, December 1988. IEEE Engineering in Medicine & Biology Society, 445 Hoes Lane, Piscataway, NJ 08854.

MEDICAL GAS ANALYZERS

Tadeusz M. Drzewiecki
Jerry M. Calkins
Defense Research Technologies, Inc.
Rockville, Maryland

INTRODUCTION

Medical gas monitoring has been so successful in improving patient safety and reducing patient risk that it has become standard medical practice today in every part of medical practice from hospital to home. The argument for providing additional patient safety will continue to be a powerful incentive to improve and enhance the methods and techniques to provide increased knowledge of the monitoring of respiratory and anesthetic gases. Research on gas markers to aid in diagnosis is an equally important application, and with the capability to measure gases at parts per billion or even trillion, heretofore unobserved gases can point to early diagnosis of such nearly always fatal neonatal diseases as necrotizing enterocolitis and other difficult-to-diagnose states.

Medical gas analyzers are used to sample and measure gases of medical importance, such as anesthesia and respiratory monitoring, and detection of trace gases for diagnostic purposes. This article predominantly discusses these two cases. The estimation of arterial blood gases is considered only in terms of measurement of respired gases. Two basic categories of sensor/analyzers exist: continuous and batch. Gas analyzers are further broken down by their sensing mechanisms into two fundamental modes of operation, specific and analytic.

Continuous devices are used where real-time information is needed. Batch systems operate on a bolus of gas, usually when real-time information is not needed. Many applications may have to live with the offline, longer duration of a batch test because nothing else is available.

Specific-type sensors rely on particular physical phenomena that are activated in the presence of a particular gas. Electrochemical devices, for example, are representative of a specific sensor wherein a voltage is developed by a chemical reaction between the sensor material and the gas being analyzed in some identified or known proportion to the amount of gas present. The same is true of fuel cells and other galvanic devices where an electric potential is developed in the presence of a difference in partial pressures across a conducting medium.

Specificity is a major issue and is of particular importance to the medical community. At this point in time, no truly specific sensors exist. All sensors exhibit some form of cross-sensitivity to a variety of gases, some in the same family, some with similar physical properties, and some for extraneous reasons not always obvious to the user. Nitrous oxide (N_2O) and carbon dioxide (CO_2) have practically the same molecular weight, 44.0128 versus 44.0098. Consequently, they have near-identical physical characteristics such as specific heat and viscosity. Interestingly, they also have almost exactly the same absorption wavelengths, although not necessarily because the atomic weights are the same but rather because the orbital electron transition energetics are similar. Thus, it is difficult to distinguish the two with most conventional techniques, and a carbon dioxide sensor that is based on absorption at 4.3 mm will be affected by the presence of nitrous oxide with its peak absorption at 4.5 mm. Unless a very narrow wavelength light source is used, such as a laser, the nitrous oxide will absorb some of the energy and make it appear that more carbon dioxide is present than there really is.

Analytic devices imply an ability to assay a gas or gas mixture and tell the user not only how much of a particular gas is present, but also which gases or elements are present and in what relative quantities, of which optical spectroscopy is a good example where a large number of absorption lines exist for different gases, so one can scan the entire spectrum from ultraviolet (UV) to far infrared (IR) and compare absorption lines to see what is present. This example is interesting because IR spectroscopy can also be the basis for a specific sensor when only a single or a particular wavelength of light is monitored looking only for a gas at that absorption line. Gas chromatography (GC) is also a batch process where a bolus of gas to be assayed is separated into its constituent parts in time by a molecular sieve and the binary pairs (the carrier and the separated gas) are detected and quantized upon exiting the GC column.

Perhaps the most common use of and need for continuous gas analysis or sensing is in real-time respiratory applications where inspired and expired (end-tidal) concentrations of respiratory gases are measured to validate that the appropriate standards of care are being applied and that proper ventilation and oxygenation of a patient is being achieved. An example would be monitoring of a ventilated patient in the ICU or recovery room. In addition, the monitoring of anesthetic gases during and immediately following anesthetic administration in the operating room is a critical application that can mean the difference between life and death. Too much can lead to brain damage or death, and too little can result in unnecessary pain and memory recall. And, of course, the detection and warning of the presence of toxic gases such as carbon monoxide released from desiccated soda lime CO_2 scrubbers due to interactions between the anesthetic agents and various scrubbers, or the production of the highly toxic Compound A, is critical to patient safety.

Continuous medical gas monitoring provides the clinician with information about the patient's physiologic status, estimates of arterial blood gases, verifies that the appropriate concentrations of delivered gases are administered, and warns of equipment failure or abnormalities in the gas delivery system. Monitors display inspired and expired gas concentrations and sound alarms to alert clinical personnel when the concentration of oxygen (O_2), carbon dioxide (CO_2), nitrous oxide (N_2O), or volatile anesthetic agent falls outside the desired set limits.

Medical gas analysis has been driven by a need for safety and patient risk reduction through respiratory gas analysis and identification and quantification of volatile anesthetic vapors. Perhaps one of the earliest anesthetic agent sensors was the Drager Narkotest, which comprised a polymer rubber membrane that contracted and moved a needle as it absorbed agent. It did not require an electrical power supply and its slow response was not any slower than the rate of change of gas composition. Much has transpired since those early days.

Currently, numerous methods and techniques of gas monitoring are in place, and new techniques and paradigms are constantly being developed to meet a new need or to serve the community with better performance or lower cost. In this review, many of the intrinsic advantages and disadvantages of these methods and techniques are discussed. A brief comparison, which includes stand-alone and multioperating room gas monitors that can determine concentrations of anesthetic and respiratory gases in the patient breathing circuit during anesthesia, is also presented.

Much of the research and development of these monitors have followed the long use of similar detector principles from analytical chemistry. As a result of the fast pace of sensor development, to a great extent driven by the need for hazardous gas sensors in the face of terrorist threats and being spearheaded by agencies such as the Defense Department's Defense Advanced Research Projects Agency (DARPA), an attempt is made to cover the most common systems and provide insights into the future based on solid technological developments.

The current development of gas analyzers is described in the extensive anesthesia and biomedical engineering literature. Complete and specific historical information about the principles and applications of these devices is well reviewed in several texts [e.g., Ref. (1)], manufacturers' and trade publications [(2) (ECRI)], and an extensive open literature describing equipment and operating principles, methods, and techniques that is available on the Internet. Societies and professional associations also exist that deal with just one method of gas analysis that can provide in-depth information about their particular interests. The Chromatographic Society is one such organization. It is the purpose of this article to concisely summarize such a large selection of information sources to a manageable few, but with enough references and pointers to allow even the casual reader to obtain whatever relevant information at whatever level is required.

CURRENT GAS MONITOR METHODS AND TECHNIQUES

As a result of the chemically diverse substances to be measured, medical gas analyzers commonly combine more than one analytical method. Methods of interest to the medical practitioner, clinician, researcher, or operator include, in alphabetical order:

- Colorimetry
- Electrochemistry
  - Fuel cells
  - Polarography
- Gas chromatography
  - Flame ionization
  - Photoionization Detectors
  - Thermal conductivity
- Infrared/Optical Spectroscopy
- Luminescence/fluorescence
- Mass spectrometry
- Nuclear Magnetic Resonance
- Paramagnetism
- Radioactive ionization
- Raman Laser Spectroscopy
- Solid-state sensors
  - Semiconductor metal oxides
  - ChemFETs
  - Solid-state galvanic cells
  - Piezoelectric/Surface Acoustic Wave

Each of these methods will be described in the following text. Illustrative examples of typical devices may be mentioned.

COLORIMETRY

Colorimetry is one of the oldest methods of gas analysis that is typically used to detect the presence of carbon dioxide in a breath as a means of determining if proper tracheal intubation has been performed. Basically, it works on the principle of changing the color of a material such as paper or cloth impregnated with a reagent in the presence of a known analyte. An example is the changes in litmus paper from purple to red in the presence of an acid and blue in the presence of a base. Carbon dioxide (CO_2) in the presence of water vapor in the exhaled breath produces carbonic acid, which turns the litmus paper toward red, usually some shade of pink. The degree of color change can be quite subjective, but a color scale usually accompanies most devices so that a coarse estimate, roughly $\pm 0.5\%$ CO_2 by volume, can be made. More expensive, sophisticated devices offer an electronic colorimetric analyzer that does the comparison automatically and can even provide a digital output.

Reagents may be tuned to a variety of specific gases and are quite commonly used in the semiconductor business to monitor levels of hydride gases to include arsine and phosphene. Hydrogen sulfide (H_2S) is also a common analyte for colorimetric sensors (1).

Cross-sensitivity can be additive or subtractive with colorimetric devices. For example, a colorimetric capnometer (CO_2 sensor) may register false-positives and false-negatives. Color may change in the presence of acidic reflux or the ingestion of acidic liquids (lemonade, wine, etc.). Conversely, the presence of bases could negate the acid response, which is true in other applications as well. For example, in hydrogen sulfide detection, the presence of methyl mercaptan (CH_4S) compounds can cancel out any reading of H_2S. Clearly, any chemical reactions between the selected analyte and a reactive contaminant can affect readings one way or another.

Figure 1 shows a photograph of one of the newer colorimetric capnometers on the market manufactured by Mercury Medical (www.mercurymed.com). A color-changing tape used in this device turns yellow in the

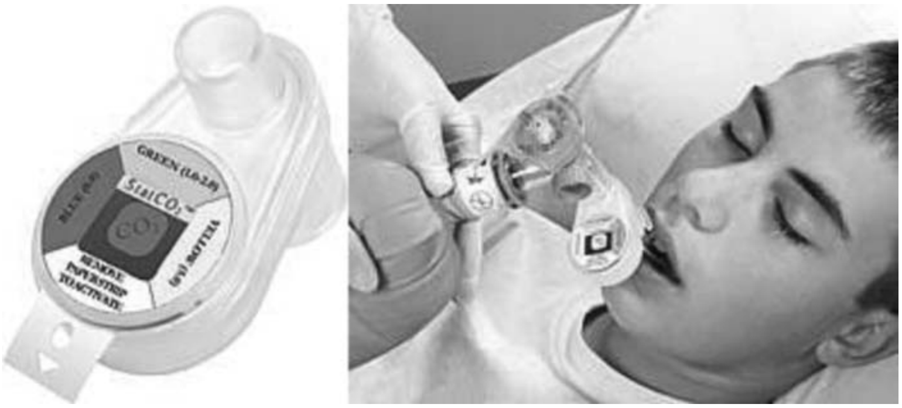


Figure 1. Mercury medical colorimetric end tidal CO_2 detector.

presence of CO_2, but returns to green when no CO_2 is present. The same piece of paper will register changes as the patient breathes. As the tape is consumed, it can be pulled through the device, and a fresh piece exposed to the breath.

ELECTROCHEMISTRY

Electrochemical gas sensors operate on the principle that a current is generated when the selected gas reacts at an electrode in the presence of an electrolyte, not unlike a battery. For this reason, these devices are often called amperometric gas sensors or microfuel cells. Electrochemical gas sensors are found ubiquitously in industry because of their excellent sensitivity to toxic gases (often low parts per million, ppm) and relatively low cost. They are, however, consumable devices and have a limited operating and shelf life, typically a few years. They are not particularly susceptible to poisoning, that is, being contaminated or degraded by absorption of particular contaminant gas species. Gases of medical importance that can be sensed with electrochemical devices are ammonia, carbon monoxide, nitric oxide, oxygen, ozone, and sulfur dioxide. The most prominent medical use is in the measurement of oxygen.

Three basic elements exist in any electrochemical gas sensor. The first is a gas-permeable, hydrophobic membrane, which allows gas to diffuse into the cell but keeps the liquid or gel electrolyte inside. It is the slow diffusion process that limits the time response of these sensors, although, if they are made very small, they can be quite responsive. Typically, however, an oxygen sensor may take as long as 20 s to equilibrate, making these devices impractical for real-time monitoring of respiration other than monitoring some average oxygen level.

The second element is the electrode. Selection of the electrode is critical to the selectivity of an appropriate reaction. Typically, electrodes are catalyzed noble metals such as gold or platinum. Gases such as oxygen, nitrogen oxides, and chlorine, which are electrochemically reducible, are sensed at the cathode while those that are electrochemically oxidizable, such as carbon monoxide, nitrogen dioxide, and hydrogen sulfide, are sensed at the anode.

The third element is the electrolyte that carries the ions between the electrodes. The electrolyte must be kept encapsulated in the cell as leakage would cause dysfunction.

In many cases, a fourth element exists which is a filter/scrubber mounted across the face of the sensor and the permeable membrane, which helps with the specificity by eliminating some interfering gases. In an oxygen sensor, this element is often an activated charcoal molecular sieve that filters out all but carbon monoxide and hydrogen. Other filters can be tailored to allow only the selected analyte through.

An oxygen fuel cell gas detector uses a lead anode that is oxidized during operation. It is powered by the oxygen it is sensing with a voltage output proportional to the oxygen concentration in the electrolyte. In this case, the electrolyte is potassium hydroxide (KOH) solution. With the recent explosion in research on fuel cells, they have become almost ubiquitous in medical practice, supplanting the Clark electrodes to a great extent.

Among the most recognizable oxygen-sensing devices are the Clark electrodes (Ag/AgCl anode, Pt cathode). One of the first applications of this device was monitoring oxygen concentrations in the inspiratory limb of the breathing circuit of an anesthesia machine. Clark electrodes differ slightly from the above-described fuel-cell-type devices. Clark electrodes require an external voltage source to generate a bias voltage against which the oxygen-induced potential operates. These devices are, therefore, called polarographic because of the bias voltage that is required for its operation, contrasting with a galvanic or amperometric cell, which produces current on its own proportional to the amount of analyte present. Oxygen from the sample fluid equilibrates across a Teflon membrane with a buffered potassium chloride (KCl) solution surrounding a glass electrode. The electrode has a platinum cathode and a silver/silver chloride anode. With between 0.5 V and 0.9 V applied across the electrodes, the consumption of O_2 at the cathode, and hence the current in the circuit, is dependent on the O_2 concentration in the solution, which rapidly equilibrates with the sample. In practice, 0.68 V is used. Performance is adversely affected by the presence of N_2O and halogenated anesthetic agents such as halothane. Protection of the platinum

Figure 2. Datex-Ohmeda Smart Vent 7900 ventilator monitor uses a galvanic cell O_2 sensor.

cathode and the need for semipermeable membranes reduces their effectiveness.

Fuel cell and polarographic devices both require temperature and pH compensation and, as indicated before, have limited life spans because of the consumable nature of the reaction and the propensity of the permeable membranes to eventually lose the effectiveness of the Clark electrodes.

Datex-Ohmeda, one of the largest manufacturers of anesthesia equipment, sells many of their systems, included with which is a galvanometric fuel cell oxygen sensor. Figure 2 shows a Smart Vent 7900$^{\mathrm{TM}}$ (3) display showing the level of oxygen being delivered. The oxygen sensor life is specified at 18 months.

GAS CHROMATOGRAPHY

Gas chromatography (GC) is an analytic tool that provides the user with an assay of what is in the gas sample of interest. It consists of two parts: (1) separation of different species by differential travel times through a separation column and (2) analytic detection of the quantity of each analyte at the end of the column using any one of a variety of methods. That is, GC provides a quantification of the constituent gases as well as an identification of what the constituent gases are. It is actually a very simple process, and, were it not for the relatively long analysis time, usually on the order of several minutes to as long as an hour, and its batch nature, it would be used more frequently.

GC requires the separation of a gas sample into its constituent gases, which is accomplished by mixing the sample with a carrier gas, usually some inert gas like helium or nitrogen, which is not adsorbed by the GC column, and passing it through a column or long impermeable, nonreacting (e.g., stainless steel) capillary filled with a zeolite, molecular sieve, or other material that separates the sample gases according to their physical or chemical properties. The different gas constituents travel through the column at different speeds and exit as binary pairs (a constituent and the carrier) in order of their adsorption properties. The time it takes for a constituent to exit the column identifies the constituent. The capillary columns can be as short as a few centimeters to tens and even hundreds of meters long. The capillary columns are heated to maintain a constant temperature, as the transport characteristics tend to be temperature-dependent.

Calibration is required to determine the transit times for each analyte, which is done by injecting known gas samples at the start of the column and physically timing them at the exit. At the exit of the column, a gas detector exists usually a flame ionization or thermal conductivity detector that measures the amount of constituent relative to the carrier. After all the constituents have been accounted for, the assay of the original sample can be made by summation of the relative amounts of each constituent and then taking the ratio of each constituent relative to the summation, which then gives the concentrations and the final assay.

Usually, this summation is accomplished by summing the areas under the detector output peaks and then ratioing the areas under the individual peaks relative to the total area. The choice of gas chromatographic detectors depends on the resolution and accuracy desired and includes (roughly, in order from most common to the least): the flame ionization detector (FID), thermal conductivity detector (TCD or hot wire detector), electron capture detector (ECD), photoionization detector (PID), flame photometric detector (FPD), thermionic detector, and a few more unusual or expensive choices like the atomic emission detector (AED) and the ozone- or fluorine-induced chemiluminescence detectors.

The Flame Ionization Detector (FID) is widely used to detect molecules with carbon-hydrogen bonds and has good sensitivity to low ppm. Basically, in operation, the analyte is injected into a hydrogen carrier and ignited inside a grounded metal chamber. Hydrogen is used because it burns clean and is carbon-free. The latter is important because output is proportional to the ionized carbon atoms. An electrode is situated just above the flame and a voltage potential is applied. The current produced is proportional to the number of carbon atoms in the analyte. When applied at the exit of a GC, very accurate measures of hydrocarbon gases can be made.

Photoionization Detectors (PIDs) are used to detect the volatile organic compound (VOC) outputs of GCs but are also widely used in industry and science to detect environmental and hazardous gases. They operate on a similar principle to that of the FID, but use ultraviolet light (UV) as opposed to a flame to ionize the flowing gas between insulated electrodes. As UV energy is a much higher frequency (lower wavelength) than IR or visible light, it can be larger and, consequently, can readily ionize gases. The ionization potentials of the analyte gases are matched by adjusting the frequency of the emitted light. The output power of the lamp is roughly the product of the number of photons and the energy per photon divided by the area and time, although changing the output frequency will change the photon energy ($E = h\upsilon$), thereby changing the power. The output power can be changed independently by increasing the fluence of photons.

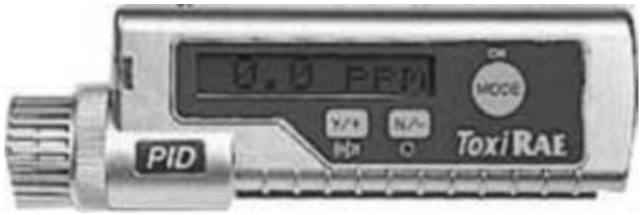


Figure 3. Seito ToxiRae personal PID gas monitor.

An inert gas lamp provides the UV light, (e.g., xenon lamps emit UV light at 147.6 nm, krypton at 123.9 nm, and argon at 105.9 nm). An advantage is that the sensitivity to particular species or groups of compounds can be adjusted by adjusting the output power to match the distinct ionization potentials of analyte gases. Consequently, different sensor sets can be achieved. For example, amines, aromatic compounds, and benzene are highly detectable at 9.5 eV. Disease and other anomaly marker gases often found in the breath, such as acetone, ammonia, and ethanol, are detectable at 9.5 eV as well as 10.6 eV. Other, more complex, marker gases such as acetylene, formaldehyde, and methanol can be detected at 10.6 eV and 11.7 eV. Typically, the PID devices are fairly responsive, on the order of a few seconds, and do well with moderately low concentrations (e.g., 0.1 ppm isobutylene).

One of the nice things about PIDs is that they can be made very small in size, as shown in Fig. 3, which shows the Rae Systems, Inc. ToxiRae personal gas monitor (www.raesystems.com). Depending on the UV source, CO, NO, SO_2, or NO_2 can be read. It works with rechargeable batteries.

Thermal Conductivity Detectors (TCD) are used to detect and quantify gases that have large variations in thermal conductivity. Gases that are discriminated well are sulfur dioxide and chlorine, which have roughly one-third the conductivity of air to helium and hydrogen, which have six and seven times the conductivity of air. As heat transfer depends on three mechanisms, radiation, convection, and conduction, the actual TCD sensor itself must be designed in such a way that conduction dominates, which implies a very slow, constant, moving flow to minimize or stabilize convection effects and a radiation-shielded enclosure. Most arrangements use two identically heated coils of wire comprising two legs of a Wheatstone bridge, one coil in a reference gas tube and the other in the sample tube. When the thermal conductivity of the sample increases, the sample coil is cooled more than the reference, and its resistance changes (usually decreases), thereby generating a voltage difference across the bridge. TCDs are usually used in high concentration applications, as they do not have the sensitivity of other techniques. TCDs do very well when mounted at the exit of a GC where the separated gas analytes are expected to have large variations in thermal conductivity.

Figure 4. An HP/Agilent laboratory-grade gas chromatograph.

Gas chromatographs have come a long way over the last decade as far as size and cost are concerned. Although laboratory-grade devices such as the HP stand-alone system shown in Fig. 4 still are fairly common, portability is being stressed in order to get the almost incomparable detectibilty of the GC to where the real gas problems exist, such as in the emergency or operating rooms, in the field, and at sites where toxins and suspected hazardous gases may be present. Bringing the GC to the patient or taking data without having subjects come into the lab has spawned systems such as Mensanna's VOC (volatile organic compound) GC system that uses a PID (Fig. 5) to check trace gases in the breath, and HP's has introduced a briefcase-sized micro-GC (Fig. 6). Lawrence Livermore National Laboratory has taken the recent developments in micromachining, MEMS (micro-electromechanical systems), and microfluidics and developed a real micro-GC. Researchers at LLNL (4) have micro-machined a very long capillary on a silicon chip, which serves as the separating column.

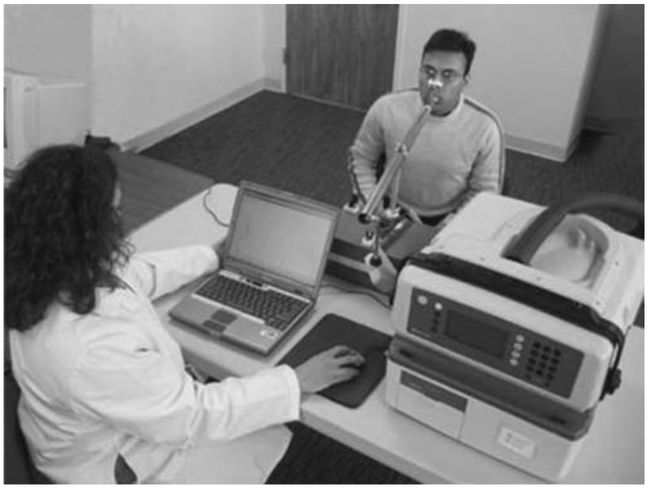


Figure 5. Mensanna portable GC for measuring breath VOCs.

Figure 6. Agilent (formerly HP) Micro-GC.

Figure 7 shows the implementation of this device that is reported to have a response time of less than 2 min.

INFRARED/OPTICAL SPECTROSCOPY

Gases absorb light or photon energy at different wavelengths depending on the complexity of their molecular structure. When a molecule is subjected to light energy of a particular frequency, the atoms involved in the atomic bonds will vibrate at the light frequency. If the frequency matches their resonant frequency, or a harmonic, they will resonate, thereby becoming highly absorbent as more of the light energy is used to feed the motion of the resonating molecules. The more complex a molecule, the greater number of different atomic bonds it will have and, consequently, the more absorption frequencies it will have. Table 1 provides some guidelines for absorption for different molecules.

Most infrared analyzers measure concentrations of volatile fluorocarbon halogenated anesthetic agents, carbon dioxide, and nitrous oxide using nondispersive infrared (NDIR) absorption technology. The transduction means may differ. Most use an electronic IR energy detector of one sort or another, such as a bolometer, solid-state photon detectors, and thermopiles; however, one monitor uses another IR detection principle, photoacoustic spectroscopy, based on the level of sound produced when an enclosed gas is exposed to pulsed/modulated IR energy.

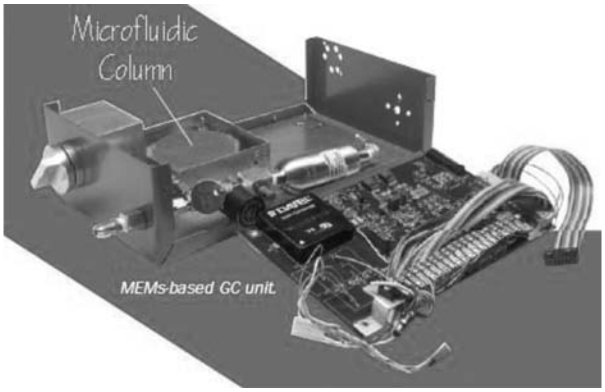


Figure 7. LLNL MEMS-based micro-GC.

Table 1. Infrared Absorption Bands for Gases of Medical Interest

Wavelength (microns)	Elements of Atomic Bonds	Typical Gases
2.7–4 μm	X-H (X = C, N, O, S)	H_2O, CH_4,
4.3–5 μm	C-X (X = C, N, O)	CO_2, N_2O
5.26–6.6 μm	C-X (X = C, N, O)	fluorocarbons
7.7–12.5 μm	C-X (X = C, N, O)	fluorocarbons

Infrared analyzers have been used for many years to identify and assay compounds for research applications. More recently, they have been adapted for respiratory monitoring of CO_2, N_2O, and halogenated anesthetic agents.

Dual-chamber NDIR spectrometers pass IR energy from an incandescent filament through the sample chamber and an identical geometry but air-filled reference chamber. Each gas absorbs light at several wavelengths, but only a single absorption wavelength is selected for each gas to determine the gas concentration. The light is filtered after it passes through the chambers, and only that wavelength selected for each gas is transmitted to the detector. The light absorption in the analysis chamber is proportional to the partial pressure (concentration) of the gas. Most manufacturers use a wavelength range around 3.3 μm, the peak wavelength at which the hydrogen-carbon bond absorbs light, to detect halogenated anesthetic hydrocarbons (halothane, enflurane, isoflurane, etc.).

In one monitor that identifies and quantifies halogenated anesthetic agents, the analyzer is a single-channel, four-wavelength IR filter photometer. Each of four filters (one for each anesthetic agent and one to provide a baseline for comparison) transmits a specific wavelength of IR energy. Each gas absorbs differently in the selected wavelength bands so that the four measurements produce a unique signature for each gas. In another monitor, potent anesthetic agents are assessed by determining their absorption at only three wavelengths of light. Normally, only one agent is present so this process reduces totagent ID. However, the use of “cocktails,” mixtures of agents, usually to reduce undesired side effects of one or another agent, require very special monitoring because of the possibility of accidental overdosing.

The Datex-Ohmeda Capnomac (www.us.datex-ohmeda.com), a multigas anesthetic agent analyzer, is based on the absorption of infrared radiation. This unit accurately analyzes breath-to-breath changes in concentrations of CO_2, NO_2, and N_2O and anesthetic vapors. It is accurate with CO_2 for up to 60 breaths/min, and 30 breaths/min for O_2 (using a slower paramagnetic sensor), but N_2O and anesthetic vapors show a decrease in accuracy at frequencies higher than 20 breaths/min. The use of narrow wave-band filters to increase specificity for CO_2 and N_2O makes the identification of the anesthetic vapors, which are measured in the same wave band more difficult. It is interesting to note that IR spectroscopy can also be used on liquids, as exemplified by the Inov 3100 near-infrared spectroscopy monitor that has been offered as a monitor for intracerebral oxygenation during anesthesia and surgery. Studies with this monitor indicate that it needs a wide optode separation and the measurements are more likely those of the external carotid flow rather than the divided internal carotid circulation (5).

A subset of NDIR is photoacoustic spectroscopy, which measures the energy produced when a gas expands by absorption of IR radiation, which is modulated at acoustic frequencies. A rotating disk with multiple concentric slotted sections between the IR source and the measurement chamber may be used tmodulate the light energy. The acoustic pressure fluctuations created occur with a frequency between 20 and 20,000 Hz, producing sound that is detected with a microphone and converted totan electrical signal. Each gas (anesthetic agent, CO_2, N_2O) exhibits this photoacoustic effect most strongly at a different wavelength. This method cannot distinguish which halogenated agent is present, however. The microphone detects the pulsating pressures from all four gases simultaneously and produces a four-component magnetic signal. A monitor using IR photoacoustic technology has been developed that can quantify all commonly respired/anesthetic gases except N_2 and water vapor. Similarly, a microphone detects the pulsating pressure changes in a paramagnetic oxygen sensor (magnetoacoustics).

The Bruel & Kjaer Multigas Monitor 1304 (6) measurements use photoacoustic spectroscopy and also incorporate a pulse oximeter. It has some advantages over the Datex Ohmeda Capnomac because it uses the same single microphone for detection of all gases, displaying gas concentration in real-time.

With the development of both fixed frequency and tunable solid-state lasers, a revolution in IR spectroscopy has occured with new technical approaches appearing every year. Tuned diode laser spectroscopy, and Laser-induced Photo Acoustic Spectroscopy (a DARPA initiative) are developments that bear close watching as they mature. The ability to produce IR energy at extremely narrow bandwidths allows discrimination of very closely related gas species such as CO_2 and N_2O. In addition, most of the volatile anesthetic agents such as halothane, desflurane, isoflurane, and sevoflurane can also be thus distinguished.

An advantage of NDIR is that the sensing mechanism does not interfere or contact the sample, thus minimal chance exists that the sample would be affected. The costs of these devices have continued to decrease, with numerous companies competing to keep the prices low and attractive. Disadvantages are the susceptibility to dirt and dust in the optical path and cross-sensitivities to interfering gases.

Companies marketing anesthesia and respiratory mechanics monitors are involved in either development or promotion of NDIR. Figure 8 shows a Datex-Ohmeda Capnomac Ultima that uses NDIR for CO_2, N_2O and anesthetic analysis, and agent identification. The top waveform is the plethysmograph, the next down is the O_2 (measured with a fast paramagnetic sensor), and the bottom waveform is the capnographic CO_2 waveform. As an added capability beyond gas analysis, to the right of the capnogram is the pressure-volume loop that is used to assess the lung compliance.

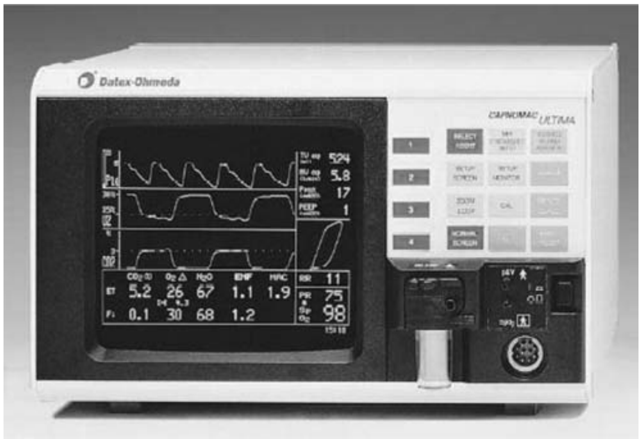


Figure 8. A Datex-Ohmeda Capomac Ultima multiple gas analyzer.

Another limitation of NDIR is its relatively low sensitivity due, for the most part, to the short path length over which the IR energy is absorbed. The short path length and small chamber size is dictated by the need for fast response in order to be able to monitor human physiological and respiratory response.

Breath rates are normally about 10 breaths per min (bpm) but, under acute hyperventilation conditions, can reach 100 bpm and higher. Also, neonates and small animals naturally exhibit high breathing rates, which requires a sensor with a millisecond response in order to be able to detect breath-by-breath variations. However, in cases where a fast response is not the driving factor, the sampling chamber may be lengthened or the path length increased.

Notwithstanding this limitation, the recent invention of Cavity Ring-Down Spectroscopy (CRDS) by Princeton chemist Kevin Lehmann (7,8) is based on absorption of laser energy over huge path lengths but is extremely fast. By bouncing laser energy between near-perfect mirrors in a sample or test chamber, the light can pass through the gas of interest multiple times, often for total distances of up to 100 km, which creates the opportunity to detect miniscule trace amounts of the gas species of interest. The time it takes for the light energy to get attenuated to zero provides a measure of the amount of the gas species present. The shorter the Ring-Down time, the more of the gas is present. These times are, however, on the order of only milliseconds. In fact, Tiger Optics of Warrington, PA, the company that has licensed Dr. Lehmann's technological development, claims that trace gases can be detected in the hundreds of parts per trillion. The LaserTrace multi-point, multi-species, multi-gas analyzer (shown in Fig. 9) is capable of detecting many species such as H_2O, O_2, CH_4, H_2, CO, NH_3, H_2S, and HF. The O_2 module measures down to 200 parts-per-trillion (ppt), in milliseconds, noninvasively and can readily be adapted to respiratory breath measurements. Methane can be detected at the parts per billion level.

TigerOptics
LASERTRACE™
MULTI-POINT, MULTI-SPECIES, MULTI-GAS ANALYZER for INERT, TOXIC, and CORROSIVE GASES

Figure 9. Tiger Optics Cavity Ring-Down Spectrometer.

Although of interest in the detection and quantification of oxygen, spectroscopy has not been widely considered for this application. However, an oxygen absorption line exist in the center of the visible spectrum at 760 nm. Often neglected because of the problems associated with the narrowness of the absorption band (0.01 nm versus 100 nm for CO_2 in the IR) as well as with spurious interference from visible light, it is nonetheless an opportunity because no interference exists from any other gases of medical interest. The development of low cost, very narrow-band lasers has resulted in the successful introduction of Laser-based Absorption Spectroscopy from Oxigraf (www.oxigraf.com). With 100 ms response and ± 0.02% resolution traces, such as that shown in Fig. 10, are possible. Cost is relatively low in comparison with other technical approaches with similar capabilities.

LUMINESCENCE/FLUORESCENCE

Gas sensors that use luminescence or fluorescence basically take advantage of the phenomenon of excitation of a molecule and the subsequent emission of radiation. Photoluminescence implies optical excitation and re-emission of light at a different, lower frequency. Chemiluminescence implies the emission of light energy as a result of a chemical reaction. In both cases, the emitted light is a function of the presence of the gas species of interest and is detected by optical means. Most industrial sensors use photomultiplier tubes to detect the light, but the needs of the medical community are being met with more compact fiber-optic systems and solid-state photodetectors.

Fluorescence quenching is a subset of luminescence-based sensors, the difference being that the presence of the analyte, rather than stimulating emission of light, actually diminishes the light output. Fundamentally, fluorescence occurs when incoming light excites an electron in a fluorescent molecule to a higher energy state, and, in turn, when the electron returns to its stable state, it releases energy in the form of light.

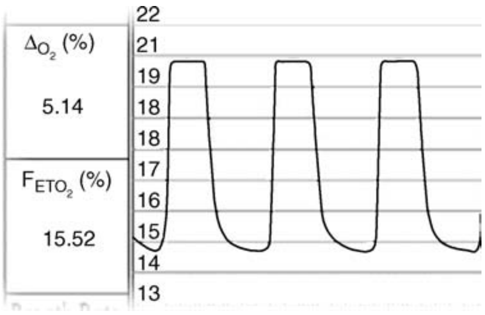


Figure 10. Oxigraf Fast Oxygen Sensor trace of respired O_2.

Two important characteristics of fluorescence are that the light necessary to excite a fluorescent molecule has a shorter wavelength than that of the fluorescent emission, and that the fluorescence of a particular molecule may be suppressed (quenched) or enhanced (dequenched) by the presence of one or more specific molecules. Consequently, the presence of such other molecules (called analytes) may be detected.

A few companies exist that use one form or another of luminescence sensing, in particular, of oxygen in the medical arena, although the sensors are ubiquitous for other gases. For example, Ocean Optics (www.oceanoptics.com) FOXY Fiber Optic Oxygen Sensors use the fluorescence of a ruthenium complex in a sol-gel to measure the partial pressure of oxygen. First, a pulsed blue LED sends light, at ~475 nm, to and through an optical fiber, which carries the light to the probe. The distal end of the probe tip consists of a thin layer of a hydrophobic sol-gel material. A ruthenium complex is trapped in the sol-gel matrix, effectively immobilized and protected from water. The light from the LED excites the ruthenium complex at the probe tip and the excited ruthenium complex fluoresces, emitting energy at ~600 nm. When the excited ruthenium complex encounters an oxygen molecule, the excess energy is transferred to the oxygen molecule in a nonradiative transfer, decreasing or quenching the fluorescence signal. The degree of quenching is a function of the level of oxygen concentration pressure in the film, which is in dynamic equilibrium with oxygen in the sample. Oxygen as a triplet molecule is able to efficiently quench the fluorescence and phosphorescence of certain luminophores. This effect is called "dynamic fluorescence quenching." When an oxygen molecule collides with a fluorophore in its excited state, a nonradiative transfer of energy occurs. The degree of fluorescence quenching relates to the frequency of collisions and, therefore, to the concentration, pressure, and temperature of the oxygen-containing media. The energy is collected by the probe and carried through an optical fiber to a spectrometer where an analog-to-digital (A/D) converter converts the data to digital data for use with a PC.

MASS SPECTROSCOPY

Mass spectroscopy provides, what many consider, the best accuracy and reliability of all of the gas analyzing/assaying schemes. The basic concept is to assay the analyte by reducing it into ionized component molecules and separating them according to their mass-to-charge ratio. By this technique, the constituents of the sample gas are ionized. The resulting ions are accelerated through an electrostatic field and then passed through a deflecting magnetic field. The lighter ions will deflect more than the heavier ions. Detecting the displacement and counting the ions can achieve the assay of the gas sample.

Ion detectors usually comprise an electric circuit where the impinging ions generate a current, which can be measured with a conventional circuit. The more current, the more ions are impinging. In practice, the ionization must be conducted in a high vacuum of the order of 10^{-6} torr. The gas is ionized with a heated filament, or by other means as have been discussed before.

A number of different mass spectroscopic configurations have been developed over the years. Time-of-flight (TOF) systems differ from the magnetically deflected devices in that the ions are free to drift across a neutrally charged evacuated flight chamber after having been accelerated electrostatically by a series of gratings that separate the ions. The time it takes for the ions to travel across the chamber is a function of their mass. An output not unlike that of a GC is developed. Quadrupole mass spectrometers focus the ions through an aperture onto a quadrupole filter. The ion-trap mass spectrometer traps ions in a small volume using three electrodes. An advantage of the ion-trap mass spectrometer over other mass spectrometers is that it has a significantly increased signal-to-noise ratio because it is able to accumulate ions in the trap. It also does not require the same kind of large dimensions that the TOF and magnetically deflected devices need, so, as a consequence, it can be made in a fairly compact size. Finally, the Fourier-transform mass spectrometer takes advantage of an ion-cyclotron resonance to select and detect ions. Single-focusing analyzers use a circular beam path of 180°, 90°, or 60°. The various forces influencing the particle separate ions with different mass-to-charge ratios. Double-focusing analyzers have an electrostatic analyzer added to separate particles with difference in kinetic energies.

A particular advantage of the mass spectrometer is that it can operate with an extremely small gas sample and can detect minute quantities. Response was long compared with most continuous sensors, but with the development of high speed microprocessors, analysis times have steadily decreased to where, today, it is not unusual to have assays in less than one minute. With the development of MEMS TOF devices, the time-of-flight is measured in microseconds.

Mass spectrometers have always tended to be bulky and expensive and, thus, rarely used on a single patient basis. Multiplexing up to 30 patients using a complex valving switching system has been shown to be feasible and has made the system much more cost-effective. Figure 11 shows a conventional ThermoElectron laboratory-grade mass spectrometer setup.

The move to miniature mass spectrometers has been rapid over the last decade, from a suitcase-sized miniature TOF mass spectrometer, developed at Johns Hopkins Applied Physics laboratory (Fig. 12) (9), to a micro-electro-mechanical system (MEMS) device smaller than a penny, developed in Korea at the MicroSystems Lab of Ajou University (Fig. 13) (10).

Mass spectroscopy is often used as the detector in combination with gas chromatography to enhance the sensitivity down to ppb, because in a GC, detection limits the capability.

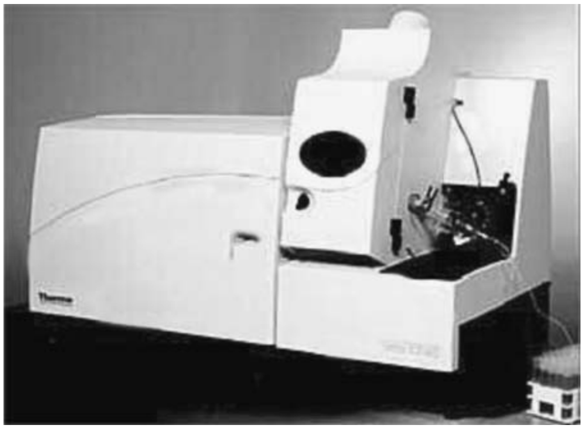


Figure 11. A ThermoElectron laboratory-grade quadrupole mass spectrometer.

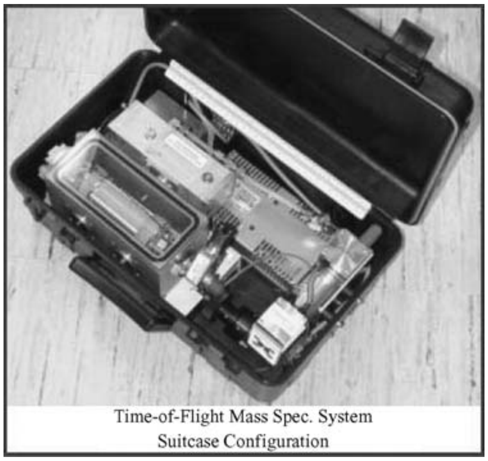


Figure 12. JHU Teeny mass spectrometer.

NUCLEAR MAGNETIC RESONANCE

Nuclear Magnetic Resonance (NMR) is the process by which a relatively low intensity radio-frequency (RF) signal at the resonant frequency of the species of gas of interest interacts with the atoms in a gas and aligns them momentarily, which requires some energy. When the RF

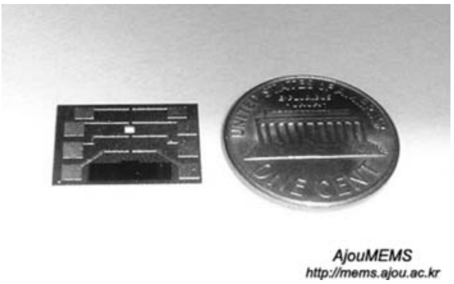


Figure 13. MEMS TOF mass spectrometers developed at Ajou University in Korea.

signal is removed, the atoms release the energy that had been stored in the alignment resonance, return to their chaotic orientations, and re-emit an RF signal at the same resonant frequency at which they were excited. Each atomic bond has its own characteristic frequency so that a spectroscopic analysis can be made by scanning through a large number of frequencies. A variant of NMR, nuclear quadrupole resonance (NQR) is used for detecting explosive vapors, which could be very useful in military medicine where explosives in clothing during triage pose a major hazard. The hydrogen bonds in TNT have a resonance at $\sim$760 kHz and the vapors of plastic explosives have resonances at low MHz frequencies. NMR is very attractive because gas analysis can be performed without having to physically take a sample of the analyte in question. As the initiating and re-emitted RF signals can pass through most nonferrous materials with little attenuation, gas, liquid, and solid chemical species can be interrogated noninvasively. By summing the returns from a number of signals, sensitivity to the low ppm can be achieved.

In dirty environments where optical or infrared devices suffer significant degradation of performance, NMR is particularly useful. Compared with chromatographic approaches, NMR eliminates the need for solvents, columns, carrier gases, or separations. Also, NMR analysis can be performed in real-time because typical atomic relaxation times, the time it takes for the atomic spin axes to return to their original orientations, is on the order of milliseconds for many gases of interest.

PARAMAGNETISM

Many gases exhibit magnetic sensitivity, paramagnetism, due to their polar nature, which means that they are attracted to a magnetic field. For oxygen, its paramagnetic sensitivity is due to its two outer electrons in unpaired orbits. Most of the gases used in anesthesia are repelled by a magnetic field (diamagnetism).

Paramagnetic sensors are typically used specifically for measuring oxygen concentration. The high degree of sensitivity of oxygen (compared with other gases) to magnetic forces reduces the cross-sensitivity to other gases of paramagnetic sensors. Sensors come in two variants, the older balance type and the newer pressure type. The balance types of sensors are relatively frail and have been replaced by the pressure types. Nevertheless, some of these older devices are still being used. The balance type of sensor uses a mechanical approach that has a dried gas sample flowing through a chamber in which a nitrogen-filled dumbbell is balanced in a magnetic field. The paramagnetic force on the oxygen in the sample puts a torque on the dumbbell. The output can be read either as a spring-loaded displacement or, in newer devices, electronically by measuring the current required to keep the dumbbell centered.

Most modern paramagnetic oxygen sensors consist of a symmetrical, two-chambered cell with identical chambers for the sample and reference gas (often air or nitrogen). These cells are joined at an interface by a responsive differential pressure transducer or microphone. Sample and reference gases are pumped through these chambers in which a strong, usually varying, magnetic field surrounding the region acts on the oxygen molecules and generates a static pressure or a time-varying (acoustic) difference between the two sides of the cell, causing the transducer to produce a DC or AC voltage proportional to the oxygen concentration. When the magnetic field is modulated at acoustic frequencies, these devices may sometimes be referred to as magnetoacoustic.

Paramagnetic oxygen analyzers are very accurate, highly sensitive, and responsive, often with a step response of 200 ms to 90% of maximum. However, they require calibration, usually with pure nitrogen and oxygen. A major drawback is that they are adversely affected by water vapor and, consequently, require a water trap incorporated into their design. The frequency response makes then useful for measurement of oxygen on a breath-by-breath basis. The Datex Ohmeda Capnomac Ultima that was previously shown in Fig. 8 uses a paramagnetic oxygen sensor, as do many of the other mainline medical gas monitor manufacturers.

RADIOACTIVE IONIZATION

The ubiquitous smoke detector found in every house, hospital, and facility has spawned detectors for other gases such as carbon monoxide, a very important marker gas in medical diagnosis. Although usually used as devices that are set to alarm when a preset level is detected, they are also used as calibrated sensors. A very low level radioactive alpha particle source (such as Americium-241) can ionize certain gases so that, in the presence of an analyte, a circuit can be completed and current caused to flow in the detector circuit.

Ionization detectors detect the presence of invisible particles (less than 0.01 micron in size) in the air. Inside the detector, a small ionization chamber exists that contains an extremely small quantity of radioactive isotope. Americium-241 emits alpha particles at a fairly constant rate. The alpha particles, which travel at an extremely high rate of speed, knock off an electron from the oxygen and nitrogen molecules in the air passing through the ionization chamber. The free electron (negative charge) is then attracted to a positively charged plate, and the positively charged oxygen or nitrogen is attracted to a negatively charged plate, which creates a very small but constant current between the plates of a detector circuit, which in itself is a gas detection mechanism much in the same way that the other ionization detectors operated. However, when particles, such as soot particles, dust, fumes, or steam, enter the ionization chamber, the current is disrupted. If the current decreases too mid, an alarm is triggered.

The disadvantage of these devices is clearly the health hazard associated with the presence of the radioactive material. However, because the detector contains only a tiny amount of radioactive material, exposure is unlikely with proper care in handling. Another disadvantage of these sensitive detectors is the false-positive alarms that can be triggered by spurious dust and other nontoxic fumes. However, the big advantage is that ionization detectors are very sensitive and, given that false alarms are tolerable, should be considered in most alarm situations.

Another form of radioactive detector is the electron capture detector, which uses a radioactive Beta emitter (electrons) to ionize some of the carrier gas and produce a current between a biased pair of electrodes. When organic molecules that contain electronegative functional groups, such as halogens, phosphorous, and nitro groups, pass by the detector, they capture some of the electrons and reduce the current measured between the electrodes.

RAMAN LASER SPECTROSCOPY

In the 1980s, Raman scattering was first heralded as an improvement to mass spectrometry (11), although some individuals had reservations (12). Although no longer manufactured but still serviced, Ohmeda Rascal II multi-gas analyzer uses a Raman scattering of laser light to identify and quantify O_2, N_2, CO_2, N_2O, and volatile anesthetic agents. It is stable and can monitor N_2 directly and CO_2 accurately for a wide range of concentrations. One of the acknowledged disadvantages is that a possibility of some destruction of volatile anesthetic agent exists during the analysis because the concentration of halothane does appear to fall when recirculated and as much as 15% must be added. Some concern exists over the reliability of the hardware, software, and laser light source (13) that is currently being addressed by others.

Raman scattering occurs when a gas sample is drawn into an analyzing chamber and is exposed to a high intensity beam from an argon laser. The laser energy is absorbed by the various molecules in the sample and are then excited into unstable vibrational or rotational energy states, which is the Raman scattering. The low intensity Raman scattered, or re-emitted, light signals are measured at right angles to the laser beam, and the spectrum of Raman scattering lines can be used to identify various types of gas molecules. Spectral analysis allows identification of known compounds by comparison with their Raman spectra. This technique is of similar accuracy to mass spectrometry.

SOLID-STATE SENSORS

At least four types of solid-state gas sensors exist: semiconductor metal oxide (SMO) sensors; chemically sensitive field effect transistors (ChemFETs); galvanic oxide sensors; and piezoelectric or surface acoustic wave (SAW) crystal sensors.

Semiconductor metal sensors are an outgrowth of the development of semiconductor devices. Early in the development of transistors and integrated circuits, it was observed that the characteristics would change in the presence of different gases. Recalling that a transistor is basically a voltage-controlled resistor, it was discovered that the p-n junction resistance was being changed by chemical reaction with the semiconductor materials (14). Commercially available Taguchi Gas Sensors (TGS) tin oxide sensors have found a niche as electronic noses. Walmsley et al. (15) used arrays of TGS sensors to develop patterns for ether and chloroform and other vapors of medical interest.

Hydrocarbons were among the first gases to be detected, and later, hydrogen sulfide was found to be detectable. Since the first tin oxide sensors appeared in the late 1960s, it has been found that by doping transition metal oxides, such as tin and aluminum, with other oxides, that as many as 150 different gases could be specifically detected (1) at ppm levels. The heated oxide adsorbs the analyte and the resistance change is a function of the concentration of the analyte. The semiconducting material is bonded or painted in a paste to a nonconducting substrate and mounted between a pair of electrodes. The substrate is heated to a temperature such that the gas being monitored reversibly changes the conductivity of the semiconducting metal oxide material. When no analyte is present, the current thinking is that oxygen molecules capture the free electrons in the semiconductor material when they are absorbing on the surface, thereby preventing the mobility of the electron flow. Analyte molecules replace the oxygen, thereby releasing the free electrons and, consequently, reducing the SMO resistance between the electrodes.

The ChemFET is derived from a field effect transistor where the normal gate metal has been replaced with a catalytic metal or chemically sensitive alloy. The gaseous analyte interacts with the gate metal and changes the FET characteristics to include gain and resistance.

Solid-state galvanic cells are based on the semiconductor galvanic properties of certain oxides or hydrides. The zirconium oxide galvanic cell oxygen sensor is probably one of the most ubiquitous sensors in daily life. It is the sensor mounted in every automobile catalytic converter to measure its effectiveness. Zirconium oxide, when heated to a temperature of about 700 °C, becomes an oxygen ion conductor, so that, in the presence of a difference in partial pressure on either side of a tube with metallized leads coming off each side, a voltage potential (Nernst voltage) is developed. These devices are commonly called fugacity sensors. As the process is reversible, a voltage applied will cause oxygen ions to flow. This process may also be applicable to the hydrogen ions in hydrides. An advantage of these oxygen sensors over other types is that no consumable exists. Hence, life is long. However, the need for heating tends to make these devices difficult to handle and they, as well as SMOs, require significant power to power the heating elements. However, because these sensors can be very small, they can have fast response times, often less than 100 ms, which makes them suitable for use for respiration monitoring. The electronics associated with the detection circuits is simple and should be very reliable.

Piezoelectric sensors use the change in a crystal vibrational frequency or propagation of surface acoustic waves to measure the concentration of a selected analyte. Most often, an analyte sample is passed through a chamber containing two piezoelectric crystals: a clean reference crystal and a second crystal that has been coated with a compound that specifically adsorbs specific analyte gases. Organophillic coatings are used for hydrocarbons such as anesthetic vapors. The resulting increase in mass changes the coated crystal's resonant frequency or the speed of propagation in direct proportion to the concentration of anesthetic gas in the sample. Using either some form of beat frequency measurement or detection of the phase shift between the two crystals, a detection circuit can generate a signal that can be processed and displayed as a concentration.

EMERGING TECHNOLOGIES—MEMS—MICROFLUIDICS–NANOTECHNOLOGY

The development of a variety of microfluidic labs-on-a-chip is leading the charge in the state-of-the-art in gas sensing. It was noted in the gas chromatography section that microfluidic channels are being used as GC columns and in the mass spectrometry section that microfluidics plays a major role in the TOF passages and chambers. The ability to miniaturize classic technologies has opened the door to mass production as well as the ability to mix-and-match sensors and technologies. The development of electronic noses that can discriminate between thousands of different chemicals and gases is driving the need to detect odors and minute quantities of dangerous or toxic gases.

Patient safety in medicine continues to be a major driver. MEMS (micro-electromechanical systems) and nanotechnology have become the enabling technologies for placing thousands of sensors in microdimensional arrays that can be placed inside a capsule and swallowed or implanted to monitor physiological and metabolic processes. IR bolometers that can sense incredibly small temperature differences as low as 0.02 °C are already a part of today's inventory (Fig. 14), and in the future, nanotechnology elements that are merely one molecule thick and dust particle-sized may provide IR spectroscopic capability in implantable or inhaled micro-packages.

A new paradigm for gas analysis that has been enabled by the development of microfluidics was originally suggested in the 1960s by Mapleson (16), who suggested measuring a physical gas property as a way of inferring binary gas mixture composition. This concept has been extended and implemented for ternary and quaternary gas mixtures with a microfluidic gas multiple gas property analyzer (17–20). Ternary mixtures are assayed with a chip that measures viscosity with a microcapillary viscometer and density with a micro-orifice densitometer. An early prototype microfluidic lab-on-a-chip showing the microcapillaries is shown in Fig. 15. By measuring properties possessed in common by all gases, such as density, viscosity, and specific heat, a single chip can be used to analyze mixtures of any four gases. The concentrations of the constituents can be determined by simultaneously solving the equations that relate the mixture properties to the concentrations, thereby determining the relative concentrations of the mixture gases required to produce the measured properties. The only limitation to such an approach is that one must know what at least all but one of the constituents are. Microfluidic property sensors such a capillary viscometers, orifice densitometers, and speed-of-sound calorimeters can provide real-time simultaneous assays of respiratory gases (O_2, CO_2, and N_2), the simultaneity of which then enables the reduction to practice of a variety of physiologic analyzers such as metabolic rate, cardiac output, and cardio-pulmonary function that had been postulated back in the 1960s (21) but never practically implemented.

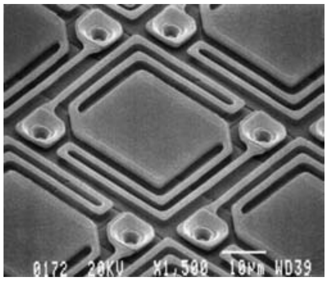


Figure 14. 25 micron wide longwave IR microbolometer detector.

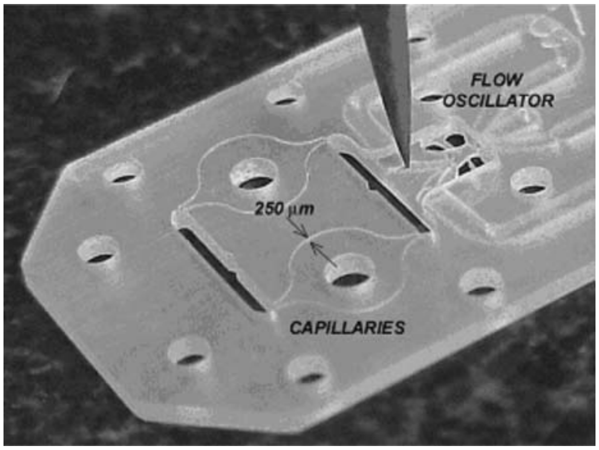


Figure 15. Microfluidic chip that measures the density and viscosity of three-component respired gases from which the constituent gas concentrations of oxygen, nitrogen, and carbon dioxide are deduced.

Advances in optics and optical fiber technology, greater expansion of solid-state sensing, and the revolutionary aspects of nanotechnology will provide gas analysis and sensing with new capabilities, provided that the lessons learned from the past are appropriately heeded.

OTHER CONSIDERATIONS IN GAS ANALYSIS

Most continuous gas analysis devices are side-stream monitors that acquire gas samples from a breathing circuit through long, narrow-diameter tubing lines. These lines may incorporate moisture removal to allow moisture to pass through the sampling line and into the atmosphere, as is the case with Nafion tubing. A water trap or filter may also be used to remove condensation from the sample in order to reduce water vapor before the gas sample reaches the analysis chamber. Gas samples are aspirated into the monitor at either an adjustable or a fixed-flow rate, typically from 50 to 250 ml/min. Lower rates minimize the amount of gas removed from the breathing circuit and, therefore, from the patient's tidal volume; however, lower sampling flow rates increase the response time and typically reduce the accuracy of conventional measurements.

Gas monitors have to eliminate the exhaust gas through a scavenging system or back to the patient's breathing circuit.

DISPLAYS, ALARMS, CALIBRATION, AND CONTROLS

Many gas monitors provide a graphic display of breath-by-breath concentrations and a hardcopy of trends of gas concentrations from the beginning to the end of an event (e.g., anesthesia delivery during an operation). The user typically calibrates or verifies calibration of the sensors with a standard gas mixture from an integral or external gas cylinder. Gas monitors are usually microprocessor-controlled and have user-adjustable alarms that typically include factory-preset default alarms or alarm-set programs for both high and low concentrations of the gases measured. Some monitors also have alarms for system malfunctions, such as disconnection from a gas source, and leaks can often be identified from trending of O_2 and CO_2. Occlusion, apnea, or inadvertent rebreathing can also be identified. Most monitors typically have a method for temporarily, but not indefinitely, silencing audible alarms for low O_2, high N_2O, and high agent, whereas other, less critical audible alarms can be permanently silenced. Most monitors typically display real-time waveforms and long-term trends. They have integral display capability and are also commonly equipped with output jacks to interface with computerized record-keeping systems or with additional analog or digital display units such as chart recorders and printers.

PATIENT SAFETY

A review of the background and significance of medical gas sensors and monitors would be incomplete without an expression of the context that patient safety has had on the impetus for recent gains in technology and the need for additional improvements. Clearly the intrinsic dangers in the conduct of anesthesia have been long understood. It became evident in the early 1980s that patient safety and reduction to risk was possible if respiratory and anesthetic gas monitoring was routinely available and used. As a result of improved and increased availability of medical gas monitoring technology and professional activity lead by the Anesthesia Patient Safety Foundation (APSF) with the support of the American Society of Anesthesiologists (ASA), a standard for monitoring has been adopted and is routinely used in today's clinical practice. This standard requires assessment of the patient's ventilation and oxygenation in addition to circulation and temperature. The use of such monitors has resulted in a significant decrease in the risk of anesthesia-related deaths and morbidity in the ICU and other critical care situations.

CONCLUSION

Medical gas monitoring has been so successful in improving patient safety and reducing patient risk that medical malpractice liability insurance companies have lowered their risk liabilities and premiums to anesthesiologists who guarantee the routine implementation of these standards whenever possible (22). The argument for providing additional patient safety will continue to be a powerful incentive to improve and enhance the methods and techniques to provide increased knowledge of the monitoring of respiratory and anesthetic gases.

The availability of gas sensors and monitors is a boon to the medical profession from both a clinical as well as a research point of view. In addition to patient safety, new diagnostic capabilities are emerging every year. In research, new gas-sensing capabilities are enhancing the discovery of markers for all kinds of disease states and metabolic functions.

Looking to the future, MEMS, microfluidics, and nanotechnology will provide growth in our understanding of physiologic processes at levels of detail never before conceived of, from inside the body as well as supplanting today's conventional techniques.

BIBLIOGRAPHY

1. Chou J, Hazardous Gas Monitors, A Practical Guide to Selection, Operation and Applicationsy. New York: McGraw-Hill; 2000.
2. ECRI. Multiple medical gas monitors, respired/anesthetic. Product Comparison System, ECRI Product Code 17–445, August 1993.
3. Datex Ohmeda Smart Vent 7900. Product Description AN2842-B/7 01 1999 Datex-Ohmeda, Inc.
4. Yu CM, Sheem SK. (1995). Miniature gas chromatography sensor. Lawrence Livermore National Lab. www.llnl.gov/sensor_technology/STR72.html.
5. Harris D, Bailey S. Near infrared spectroscopy in adults. Anaesthesia 1993;48:694–696.
6. McPeak HB, Palayiwa E, Robinson GC, Sykes MK. An evaluation of the Bruel and Kjaer monitor 1304. Anaesthesia 1992; 47(1):41–47.
7. Lehmann KK, Rabinowitz P. High-finesse optical resonator for cavity ring-down spectroscopy based upon Brewster's angle prism retroreflectors. U.S. Patent 5,973,864, October 26, 1999.
8. Lehmann KK, Romanini D. The superposition principle and cavity ring-down spectroscopy. J Chem Phys 1996;105(23): 10263–10277.
9. Ecelberger SA, Cornish TJ, Bryden W. The improved teeny-TOF mass spectrometer for chemical and biological sensing, 3rd Harsh-Environment Mass Spectrometry Workshop, March 25–28, 2002 Pasadena, CA.
10. Yoon HY, Kim JH, Choi ES, Yang SS, Jung KW. Fabrication of a novel micro time-of-flight mass spectrometer. Sens Actuators A 2002;97–98:441–447.
11. Westenskow DR, Smith KW, Coleman DL, et al. Clinical evaluation of a Raman scattering multiple gas analyzer. Anesthesiology 1989;70:350–355.
12. Severinghaus JW, Ozanne GM. Multi-operating room monitoring with one mass spectrometer. Acta Anaesthesiol Scan 1987; (Suppl) 70:186–187.
13. Lockwood G, Landon MJ, Chakrabarti MK, Whitwam JG. The Ohmeda Rascal II. Anaesthesia 1994;49:44–53.
14. Kress-Rogers E, ed. Handbook of Biosensors and Electronic Noses—Medicine, Food and the Environment. Boca Raton, FL: CRC Press; 1996.

15. Walmsley AD, Haswell SJ, Metcalfe E. Methodology for the selection of suitable sensors for incorporation into a gas sensor array. Analytica Chimica Acta 1991;242:31.
16. Mapleson WW. Physical methods of gas analysis. Brit J Anaesth 1962;34:631.
17. Drzewiecki TM. Fluidic multiple medical gas monitor. NIH BioEngineering Symposium, Bethesda, MD, February 1998.
18. Drzewiecki TM, Polcha M, Koser M, Calkins J. A novel inexpensive respiratory and anesthetic gas monitor. Society for Technology in Anesthesia 1999 Annual Meeting. San Diego, CA: January 1999.
19. Drzewiecki TM, Calkins J. Real time, simultaneous analysis of multiple gas mixtures using microfluidic technology. Proc Instrument Society of America AD 2000 Symposium, Charleston, WV: April 2000.
20. Drzewiecki TM. Method and apparatus for real time gas analysis. U.S. Patent 6,076,392, June 2000.
21. Kim TS, Rahn H, Farhi LE. Estimation of true venous and arterial PCO_2 by gas analysis of a single breath. J Appl Physiol 1966;21(4):1338–1344.
22. Swedlow DB. Respiratory gas monitoring. In: Saidman L, Smith N, eds. Monitoring in Anesthesia. 3rd ed. Boston, MA: Butterworth-Heinemann; 1993. pp 27–50.

See also BLOOD GAS MEASUREMENTS; RESPIRATORY MECHANICS AND GAS EXCHANGE.

MEDICAL PHOTOGRAPHY. See PHOTOGRAPHY, MEDICAL.

MEDICAL PHYSICS LITERATURE

COLIN ORTON
Harper Hospital and Wayne
State University
Detroit, Michigan

INTRODUCTION

Medical physicists are responsible for the application of physics to the diagnosis and treatment of disease and other disabilities although, in some countries, the treatment of patients with disabilities is a separate field, often referred to as "biomedical engineering" or words to that effect. Here, we restrict ourselves to applications in the diagnosis and treatment of disease.

The major applications in diagnosis are the use of X-rays for imaging (diagnostic radiology, including computerized tomography (CT), etc.); radioactive isotopes for imaging and uptake measurements [nuclear medicine, single photon emission computed tomography (SPECT), positron emission tomography (PET), etc.]; magnetic resonance magnetic resonance imaging (MRI) and spectroscopy (MRS); ultrasound (ultrasonography).

Applications of physics to the treatment of disease include the following: external beams of X-rays, gamma-rays, or electrons for the treatment of cancer radiation oncology, including stereotactic radiosurgery, intensity modulated radiation therapy (IMRT), total body irradiation (TBI), etc.; external beams of heavy particles for the treatment of cancer (neutrons, protons, heavy ions); internal radioisotope treatments for cancer (brachytherapy, systemic radiotherapy, and radioimmunotherapy) and other problems (intravascular brachytherapy, hyperthyroidism, etc.); hyperthermia for the treatment of cancer.

Other topics of major interest for medical physicists include various medical applications of light and lasers, radiation protection, radiation measurements, radiation biology, and the applications of computers to all of the above.

Throughout the world there are medical physics organizations that represent medical physicists and provide information to help them in their profession. Most of these are national associations, with about 70 of these represented by the International Organization for Medical Physics (IOMP). In terms of publications, by far the two most prolific organizations are the Institute of Physics and Engineering in Medicine (IPEM) in the United Kingdom, and the American Association of Physicists in Medicine (AAPM) in North America. These two organizations between them have published hundreds of reports, monographs, and meeting proceedings that are used as reference materials throughout the world. From the IPEM many of these are published by the Institute of Physics Publishing (IOPP) in Bristol, UK, and from the AAPM many are published by either the American Institute of Physics (AIP) or Medical Physics Publishing (Madison, WI).

JOURNALS

Several national and international organizations and independent publishers publish journals used by medical physicists. Some of these journals are used extensively for medical physics papers in which at least 25% of the manuscripts are medical physics articles. These are categorized as "Primary" journals below. Others contain some medical physics articles (<25%) and are categorized as "Secondary".

PRIMARY MEDICAL PHYSICS JOURNALS

Australasian Physical & Engineering Sciences in Medicine, Australasian College of Physical Scientists and Engineers in Medicine and the College of Biomedical Engineers.

Journal of Applied Clinical Medical Physics, American College of Medical Physics (http://www.jacmp.org).

Journal of Medical Physics, Association of Medical Physicists of India.

Medical Dosimetry, American Association of Medical Dosimetrists (Elsevier).

Medical Physics, American Association of Physicists in Medicine (AIP).

Physica Medica, Istituti Editoriali e Poligrafici Internazionali Casella Postale n.1, Succursale n.8, 56123 Pisa, Italy (http://www.iepi.it).

Physics in Medicine and Biology, Institute of Physics and Engineering in Medicine (IOPP).

Polish Journal of Medical Physics and Engineering, Polish Society of Medical Physics.

Zeitschrift für Medizinische Physik, Deutschen, Österreichischen und Schweizerischen Gesellschaft für Medizinische Physik (Elsevier).

SECONDARY MEDICAL PHYSICS JOURNALS

Acta Radiologica, Scandinavian Society of Radiology (Taylor & Francis).

American Journal of Roentgenology, American Roentgen Ray Society.

Applied Radiation and Isotopes (Elsevier).

Australasian Radiology, Royal Australian and New Zealand College of Radiologists (Blackwell).

Biomedical Imaging and Interventional Journal (University of Malaya: http://www.biij.org).

Biomedical Instrumentation & Technology, Association for the Advancement of Medical Instrumentation.

Brachytherapy, American Brachytherapy Society (Elsevier).

British Journal of Radiology, British Institute of Radiology.

Canadian Association of Radiologists Journal, Canadian Association of Radiologists.

Cancer Radiothérapie, Société Française de Radiothérapie Oncologique (Elsevier).

Cardiovascular and Interventional Radiology, Cardiovascular and Interventional Radiological Society of Europe, Japanese Society of Angiography and Interventional Radiology, and British Society of Interventional Radiology (Springer).

Cardiovascular Radiation Medicine (Elsevier).

Clinical Imaging (Elsevier).

Clinical Radiology, Royal College of Radiologists (Elsevier).

Critical Reviews in Computed Tomography (Taylor & Francis).

Computerized Medical Imaging and Graphics, Computerized Medical Imaging Society (Elsevier).

Computers in Biology and Medicine (Elsevier).

Current Problems in Diagnostic Radiology (Mosby).

European Journal of Nuclear Medicine and Molecular Imaging, European Association of Nuclear Medicine (Springer).

European Journal of Radiology (Elsevier).

European Journal of Ultrasound, European Federation of Societies for Ultrasound in Medicine and Biology (Elsevier).

European Radiology, European Congress of Radiology (Springer).

Health Physics, Health Physics Society (Lippincott Williams & Wilkins).

IEEE Transactions on Medical Imaging, IEEE.

International Journal of Radiation Biology (Taylor & Francis).

International Journal of Radiation Oncology, Biology, Physics, American Society of Therapeutic Radiology and Oncology (Elsevier).

Investigative Radiology (Lippincott Williams & Wilkins).

Journal of Biomedical Optics, International Society for Optical Engineering (SPIE).

Journal of Cardiovascular Magnetic Resonance, Society for Cardiovascular Magnetic Resonance (Taylor & Francis).

Journal of Clinical Ultrasound (Wiley).

Journal of Computer Assisted Tomography (Lippincott Williams & Wilkins).

Journal of Diagnostic Radiography and Imaging, Royal Society of Medicine.

Journal of Digital Imaging, Society for Computer Applications in Radiology (Springer).

Journal of Electronic Imaging, International Society for Optical Engineering (SPIE).

Journal of Labelled Compounds and Radiopharmaceuticals (Wiley).

Journal of Magnetic Resonance Imaging, International Society for Magnetic Resonance Medicine (Wiley)

Journal of Neuroimaging, American Society of Neuroimaging (Sage Publications).

Journal of Nuclear Cardiology, American Society of Nuclear Cardiology (Elsevier).

Journal of Nuclear Medicine, Society of Nuclear Medicine.

Journal of Nuclear Medicine Technology, Society of Nuclear Medicine.

Journal of Radiological Protection, Society for Radiological Protection (IOPP, Bristol, U.K.).

Journal of the Acoustical Society of America, Acoustical Society of America (American Institute of Physics, New York).

Journal of the American Society of Echocardiography, American Society of Echocardiography (Mosby).

Journal of Thoracic Imaging, Society of Thoracic Radiology (Lippincott Williams & Wilkins).

Journal of Ultrasound in Medicine, American Institute of Ultrasound in Medicine.

Journal of Vascular and Interventional Radiology, Society of Interventional Radiology (Lippincott Williams & Wilkins).

Journal of X-Ray Science and Technology (IOS Press, Amsterdam).

Lasers in Medical Science (Springer).

Lasers in Surgery and Medicine, American Society for Laser Medicine and Surgery (Wiley).

Medical Engineering & Physics (Elsevier).

Magnetic Resonance Imaging (Elsevier).

Magnetic Resonance in Medicine, International Society for Magnetic Resonance in Medicine (Wiley).

Medical Engineering & Physics, Institute of Physics and Engineering in Medicine (Elsevier).

Medical Image Analysis (Elsevier).

Molecular Imaging and Biology, Academy of Molecular Imaging (Springer).

Neuroradiology, European Society of Neuroradiology (Springer).

NMR in Biomedicine (Wiley).

Nuclear Medicine and Biology, Society of Radiopharmaceutical Sciences (Elsevier).

Pediatric Radiology, European Society of Pediatric Radiology, Society for Pediatric Radiology, Asian and Oceanic Society for Pediatric Radiology (Springer).

Photomedicine and Laser Surgery, World Association for Laser Therapy (Mary Ann Liebert, Inc.).

Physiological Measurement, Institute of Physics and Engineering in Medicine (IOPP).

Progress in Nuclear Magnetic Resonance Spectroscopy (Elsevier).

Radiation Measurements (Elsevier).

Radiation Physics and Chemistry (Elsevier).

Radiation Research, Radiation Research Society.

Radiographics, Radiological Society of North America.

Radiography, College of Radiographers (Elsevier).

Radiology, Radiological Society of North America.

Radiotherapy and Oncology, European Society for Therapeutic Radiology and Oncology (Elsevier).

Seminars in Interventional Radiology (Thieme).

Seminars in Nuclear Medicine (Elsevier).

Seminars in Radiation Oncology (Saunders).

Seminars in Roentgenology (Elsevier).

Seminars in Ultrasound, CT and MRI (Elsevier).

Techniques in Vascular and Interventional Radiology (Elsevier).

Topics in Magnetic Resonance Imaging (Lippincott Williams & Wilkins).

Ultrasound in Medicine & Biology, World Federation for Ultrasound in Medicine and Biology (Elsevier).

Ultrasound in Obstetrics and Gynecology (Wiley).

Year Book of Diagnostic Radiology (Elsevier).

Year Book of Nuclear Medicine (Elsevier).

BOOKS AND REPORTS

As with journals, books and reports are published by both medical physics organizations, especially the AAPM and the IPEM, and independent publishers. Most of the books and reports in the following lists can be purchased directly from the publishers or, alternatively, through bookstores using the ISBN number provided.

MEDICAL AND RADIOLOGICAL PHYSICS

General

A Century of X-Rays and Radioactivity in Medicine: With Emphasis on Photographic Records of the Early Years, Richard F. Mould, ISBN 0-7503-0224-0, 1993, 236 pp, IOPP, Bristol (UK).

Essentials of Radiology Physics, Charles A. Kelsey, ISBN: 0875273548, 1985, 467 pp, W.H. Green.

Introduction to Radiological Physics and Radiation Dosimetry, Frank Herbert Attix, ISBN: 0-471-01146-0, 1986, 640 pp, Advanced Medical Publishing, Madison (WI).

Meandering in Medical Physics: A Personal Account of Hospital Physics, J.E. Roberts, N.G. Trott, ISBN: 0750304944, 1999, 181 pp, IOPP, Bristol (UK).

Medical Physics and Biomedical Engineering, Brown BH, Smallwood RH, Barber DC, Lawford PV; Hose DR, ISBN: 0750303670, 1998, 768 pp, IOPP, Bristol (UK).

Medical Physics Handbook of Units and Measures, Freim J, Jr, ISBN: 0944838308, 1992, 47 pp, Medical Physics Publishing, Madison (WI).

Medical Radiation Physics: Roentgenology, Nuclear Medicine & Ultrasound, Hendee WR, ISBN: 0815142404, 1979, 517 pp, Year Book Medical Publishers.

Physics in Medicine and Biology Encyclopedia (2 Volume Set), T.F. McAinsh, (editor), ISBN: 0080264972, 1986, Pergamon, Elmsfood (NY).

Physics and Engineering in Medicine in the New Millennium, Sharp PF, Perkins AC editors., ISBN: 0904181952, 2000, 156 pp, IPEM, York, (UK).

Physics of Radiology, 4th ed., Johns H E, John Robert Cunningham, ISBN: 0398046697, 1983, 796 pp, Charles C. Thomas.

Physics of Radiology, second edition, Wolbarst A B, ISBN: 1-930524-22-6. Published: 2005, 660 pp, Medical Physics Publishing, Madison (WI).

Physics of the Body, Cameron JR, Skofronick JG, Grant RM, ISBN: 094483891X, 1999, 394 pp, Medical Physics Publishing, Madison (WI).

Principles and Practice of Clinical Physics & Dosimetry, Michael L.F. Lim, ISBN: 1-883526-11-6, 2005, 500 pp, Advanced Medical Publishing, Madison (WI).

Principles of Radiological Physics, 4th ed., Graham D, Cloke P, ISBN: 0443070733, 2003, 576 pp, Churchill Livingstone.

Radiation Biophysics, Alpen EL, ISBN: 0120530856, 1998, 484 pp, Academic Press.

Radiation Physics Handbook for Medical Physicists, Ervin B. Podgorsak, ISBN: 3540250417, 2005, 360 pp, Springer, New York.

Review of Radiological Physics, Walter Huda, Richard M. Slone, ISBN: 0781736757, 2002, 350 pp, Lippincott Williams & Wilkins, Philadelphia.

Topical

How the Body Works, Lenihan J, ISBN: 0944838-48-0, 1995, 200 pp, Medical Physics Publishing, Madison (WI).

Physics of the Body 2nd ed., Cameron J, et al., ISBN: 0-944838-91-X, 1999, 394 pp, Medical Physics Publishing, Madison (WI).

Medical Applications of Nuclear Physics, Bethge K, Kraft G, Kreisler P, Walter G, ISBN: 3540208054, 2004, 208 pp, Springer.

Progress in Medical Radiation Physics, Orton CG, ISBN: 0306417898, 1985, 248 pp, Plenum, New York.

RADIATION ONCOLOGY PHYSICS

General

AAPM Monograph No. 15, Radiation Oncology Physics, Kereiakes J, Elson H, Born C, editors, ISBN: 0-883185-33-4, 1986, 812 pp, Medical Physics Publishing, Madison (WI).

AAPM Monograph No. 26, General Practice of Radiation Oncology Physics in the 21st Century, Almon Shiu & David Mellenberg, ISBN: 0-944838-98-7, 2000, 368 pp, Medical Physics Publishing, Madison (WI).

Applied Physics for Radiation Oncology, Robert Stanton & Donna Stinson ISBN: 0-944838-60-X, 1996, 375 pp, Medical Physics Publishing, Madison (WI).

Biomedical Particle Accelerators, Scharf WH, Siebers JV, ISBN: 1563960893, 1994, 480 pp, Springer.

Blackburn's Introduction to Clinical Radiation Therapy Physics, Benjamin Blackburn ISBN: 0-944838-06-5, 1989, 218 pp, Medical Physics Publishing, Madison (WI).

Clinical Radiotherapy Physics, 2nd ed., Jayaraman S, Lanzl LH, Lanzl EF, ISBN: 3540402845, 2004, 523 pp, Springer.

Handbook of Radiotherapy Physics: Theory and Practice, Mayles P, Nahum A, Rosenwald J-C, ISBN: 0750308605, 2005, 700 pp, IOPP, Bristol (UK).

Modern Technology of Radiation Oncology, Van Dyk J, ISBN: 0-944838-38-3, 1999, 1072 pp, Medical Physics Publishing, Madison (WI).

Practical Radiotherapy: Physics and Equipment, Cherry P, Duxbury A, ISBN: 1900151065, 1998, 224 pp, Cambridge University Press, New York.

Radiation Therapy Physics, Hendee WR, Ibbott GS, Hendee EG, ISBN: 0471394939, 2004, 450 pp, Wiley, New York.

Radiotherapy Physics and Equipment, Morris S, Williams A, ISBN: 0443062110, 2001, 176 pp, Churchill Livingstone.

Radiotherapy Physics: In Practice, Williams JR, Thwaites DI, editors, ISBN: 0-19-262878-X, 2000, 362 pp, Oxford University Press, New York.

Review of Radiation Oncology Physics, Prasad SC, ISBN: 1-930524-08-0, 2002, 95 pp, Medical Physics Publishing, Madison (WI).

Study Guide for Radiation Oncology Physics Board Exams, Berman B, ISBN: 0-944838-94-4, 2000, 112 pp, Medical Physics Publishing, Madison (WI).

The Physics of Radiation Therapy, Hardbound 3rd ed., Khan F, ISBN: 0-7817-3065-1, 2003, 511 pp, Wiley, New York.

Walter & Miller's Textbook of Radiotherapy Radiation Physics, Therapy and Oncology, 6th ed., Bomford CK et al., ISBN: 0443062013, 2003, 660 pp, Elsevier.

Topical

3-D Conformal and Intensity Modulated Radiation Therapy: Physics and Clinical Applications, Purdy JA, Grant W III, Palta JR, Butler EB, Perez CA, editors, ISBN: 1-883526-10-8, 2001, 650 pp, Advanced Medical Publishing, Madison (WI).

A Practical Guide to 3-D Planning and Conformal Radiation Therapy, Purdy JA, Starkschall G, ISBN: 1-883526-07, 1999, 400 pp, Advanced Medical Publishing, Madison (WI).

A Practical Guide to Intensity-Modulated Radiation Therapy, Memorial Sloan-Kettering Cancer Center, ISBN: 1-930524-13-7, 2003, 450 pp. Medical Physics Publishing, Madison (WI).

AAPM Manual No. 2: Workbook on Dosimetry and Treatment Planning for Radiation Oncology Residents, Wu RK, Gerbi BJ, Doppke KP, editors, ISBN: 0-88318-916-X, 1991, 32 pp, Medical Physics Publishing, Madison (WI).

AAPM Monograph No. 2, Practical Aspects of Electron Beam Treatment Planning, Orton CG, Bagne F, editors, ISBN: 0-88318-247-5, 1978, 109 pp, Medical Physics Publishing, Madison (WI).

AAPM Monograph No. 7, Recent Advances in Brachytherapy Physics, Shearer DR, editors, ISBN: 0-88318-285-8, 1981, 202 pp, Medical Physics Publishing, Madison (WI).

AAPM Monograph No. 8, Physical Aspects of Hyperthermia, Nussbaum GH, editors, ISBN: 0-88318-414-1, 1982, 656 pp, Medical Physics Publishing, Madison (WI).

AAPM Monograph No. 9, Advances in Radiation Therapy Treatment Planning, Wright AE, Boyer A, editors, ISBN: 0-883184-23-0, 1982, 626 pp, Medical Physics Publishing, Madison (WI).

AAPM Monograph No. 16, Biological, Physical and Clinical Aspects of Hyperthermia, Paliwal BR, Hetzel FW, Dewhirst M, editors, ISBN: 0-88318-558-X, 1988, 483 pp, Medical Physics Publishing, Madison (WI).

AAPM Monograph No. 28, Intravascular Brachytherapy/Fluoroscopically Guided Interventions, Balter S, Chan RC, Shope TB Jr, editors, ISBN: 1-930524-10-2, 2002, 930 pp, Medical Physics Publishing, Madison (WI).

AAPM Monograph No. 29, Intensity-Modulated Radiation Therapy: The State of the Art, Palta JR, Rockwell Mackie T, editors, ISBN: 1-930524-16-1, 2003, 904 pp. Medical Physics Publishing, Madison (WI).

AAPM Proceedings No. 2, Proceedings of the Symposium on Electron Dosimetry and Arc Therapy, Paliwal BR, editor, ISBN: 0-88318-404-4, 1981, 384 pp, Medical Physics Publishing, Madison (WI).

AAPM Proceedings No. 3, Proceedings of a Symposium on Quality Assurance of Radiotherapy Equipment, Starkschall G, editor, ISBN: 0-88318-422-2, 1982, 242 pp, Medical Physics Publishing, Madison (WI).

AAPM Proceedings No. 5, Optimization of Cancer Radiotherapy, Paliwal BR, Herbert DE, Orton CG, editors, ISBN: 0-88318-483-4, 1984, 556 pp, Medical Physics Publishing, Madison (WI).

AAPM Proceedings No. 12, Biological & Physical Basis of IMRT & Tomotherapy, Paliwal BR, Herbert DE, Fowler JF, Mehta M, editors, ISBN: 1-930524-11-0, 2002, 390 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 7, Protocol for Neutron Beam Dosimetry, Radiation Therapy Committee Task Group 18; ISBN: 0-88318-276-9, 1980, 51 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 13, Physical Aspects of Quality Assurance in Radiation Therapy, Radiation Therapy Committee Task Group 24, with contribution from Task Group 22, ISBN: 0-88318-457-5, 1984, 63 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 17, The Physical Aspects of Total and a Half Body Photon Irradiation, Radiation Therapy Committee Task Group 29; ISBN: 0-88318-513-X, 1986, 55 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 21, Specification of Brachytherapy Source Strength, Radiation Therapy Committee Task Group 32; ISBN: 0-88318-545-8, 1987, 21 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 23, Total Skin Electron Therapy: Technique and Dosimetry Radiation Therapy Committee Task Group 30; ISBN: 0-88318-556-3, 1987, 55 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 24, Radiotherapy Portal Imaging Quality, Radiation Therapy Committee Task Group 28; ISBN: 0-88318-557-1, 1987, 29 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 26, Performance Evaluation of Hyperthermia Equipment, Hyperthermia Committee Task Group 1; ISBN: 0-88318-636-5, 1989, 46 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 27, Hyperthermia Treatment Planning, Hyperthermia Committee Task Group 2; ISBN: 0-88318-643-8, 1989, 57 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 32, Clinical Electron-Beam Dosimetry, Radiation Therapy Committee Task Group 25, ISBN: 0-88318-905-4, 1990, 40 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 40, Radiolabeled Antibody Tumor Dosimetry (Reprinted from Medical Physics, Vol. 20, Issue 2), Nuclear Medicine Committee Task Group 2; ISBN: 1-56396-233-0, 1993, 112 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 41, Remote Afterloading Technology, Remote Afterloading Technology Task Group 41; ISBN: 1-56396-240-3, 1993, 107 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 45, Management of Radiation Oncology Patients with Implanted Cardiac Pacemakers (Reprinted from Medical Physics, Vol. 21, Issue 1), Task Group 34. ISBN: 1-56396-380-9, 1994, 6 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 46, Comprehensive QA for Radiation Oncology (Reprinted from Medical Physics, Vol. 21, Issue 4), Radiation Therapy Committee Task Group 40; ISBN: 1-56396-401-5, 1994, 37 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 47, AAPM Code of Practice for Radiotherapy Accelerators (Reprinted from Medical Physics, Vol. 21, Issue 4), Radiation Therapy Task Group 45; ISBN: 1-56396-402-3, 1994, 37 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 48, The Calibration and Use of Plane-Parallel Ionization Chambers for Dosimetry of Electron Beams (Reprinted from Medical Physics, Vol. 21, Issue 8) Radiation Therapy Committee Task Group 39; ISBN: 1-56396-461-9, 1994, 10 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 50, Fetal Dose from Radiotherapy with Photon Beams (Reprinted from Medical Physics, Vol. 22, Issue 1), Radiation Therapy Committee Task Group 36; ISBN: 1-56396-453-8, 1995, 20 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 54, Stereotactic Radiosurgery, Radiation Therapy Committee Task Group 42; ISBN: 1-56396-497-X, 1995, 100 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 55, Radiation Treatment Planning Dosimetry Verification, AAPM ISBN: 1-56396-534-8, 1995, 200 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 56, Medical Accelerator Safety Considerations (Reprinted from Medical Physics, Vol. 20, Issue 4), Radiation Therapy Committee Task Group 35; ISBN: 1-888340-01-0, 1993, 15 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 59, Code of Practice for Brachytherapy Physics (Reprinted from Medical Physics, Vol. 24, Issue 10), Radiation Therapy Committee Task Group 56; ISBN: 1-888340-14-2, 1997, 42 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 61, High Dose-Rate Brachytherapy Treatment Delivery (Reprinted from Medical Physics, Vol. 25, Issue 4), Radiation Therapy Committee Task Group 59; ISBN: 1-888340-17-7, 1998, 29 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 62, Quality Assurance for Clinical Radiotherapy Treatment Planning (Reprinted from Medical Physics, Vol. 25, Issue 10), Radiation Therapy Committee Task Group 53; ISBN: 1-888-340-18-5, 1998, 57 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 66, Intravascular Brachytherapy Physics (Reprinted from Medical Physics, Vol. 26, Issue 2), Radiation Therapy Committee Task Group 60; ISBN: 1-888340-23-1, 1999, 34 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 67, Protocol for Clinical Reference Dosimetry of High-Energy Photon and Electron Beams (Reprinted from Medical Physics, Vol. 26, Issue 9) Task Group 51; ISBN: 1-888340-25-8, 1999, 24 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 68, Permanent Prostate Seed Implant Brachytherapy (Reprinted from Medical Physics, Vol. 26, Issue 10), Task Group 64; ISBN: 1-888340-26-6, 1999, 23 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 69, Recommendations of the AAPM on ^{103}Pd Interstitial Source Calibration and Dosimetry: Implications for Dose Specification and Prescription, AAPM, ISBN: 1-888340-27-4, 2000, 9 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 71, A Primer for Radioimmunotherapy and Radionuclide Therapy Task Group 7, ISBN: 1-888340-29-0, 2001, 73 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 72, Basic Applications of Multileaf Collimators, Task Group 50 ISBN: 1-888340-30-4, 2001, 54 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 75, Clinical Use of Electronic Portal Imaging, Radiation Therapy Committee Task Group 58, ISBN: 1-888340-34-7, 2001, 26 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 76, AAPM Protocol for 40-300 kV X-ray Beam Dosimetry in Radiotherapy and Radiobiology, AAPM, ISBN: 1-888340-35-5, 2001, 26 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 81, Dosimetric Considerations for Patients with Hip Prostheses Undergoing Pelvic Irradiation, Radiation Therapy Committee Task Group 63 (Reprinted from Medical Physics, Vol. 30, Issue 6), ISBN: 1-888340-42-8, 2003, 21 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 82, Guidance Document on Delivery, Treatment Planning, and Clinical Implementation of IMRT, AAPM Radiation Therapy Committee, ISBN: 1-888340-43-6, 2003, 25 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 83, Quality Assurance for Computed-Tomography Simulators and the Computed-Tomography-Simulation Process, Radiation Therapy Committee Task Group 66, ISBN: 1-888340-44-4, 2003, 31 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 84, A Revised AAPM Protocol for Brachytherapy Dose Calculations (Reprinted from Medical Physics, Vol. 31, Issue 3, pp. 633-674), Radiation Therapy Committee, ISBN: 1-888340-46-0, 2004, 42 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 85, Tissue Inhomogeneity Corrections for Megavoltage Photon Beams, Task Group No. 65 of the Radiation Therapy Committee, ISBN: 1-888340-47-9, 2004, 124 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 86, Quality Assurance for Clinical Trials: A Primer for Physicists, Subcommittee on Quality Assurance Physics for Cooperative Trials of the Radiation Therapy Committee, ISBN: 1-88340-48-7, 2004, 68 pp, Medical Physics Publishing, Madison (WI).

Achieving Quality in Brachytherapy, BR Thomadsen, ISBN: 0750305541, 1999, pp, 268, Advanced Medical Publishing, Madison (WI).

A Practical Guide to CT-Simulation, Edited by: Coia L, Schultheiss T, Hanks G, ISBN: 1-883526-04-3, 1995, 216 pp, Advanced Medical Publishing, Madison (WI).

Brachytherapy Physics (1994 AAPM Summer School), Williamson J et al., ISBN: 0-944838-50-2, 1995, 715 pp, Medical Physics Publishing, Madison (WI).

Clinical Target Volumes in Conformal and Intensity Modulated Radiation Therapy, Gregorie V, Scalliet and P, Ang KK, ISBN: 3540413804, 2003, 300 pp, Springer Verlag.

Contemporary IMRT: Developing Physics and Clinical Implementation, Webb S, ISBN: 0750310049, 2004, 478 pp, IOPP, Bristol (UK).

CT Simulation for Radiotherapy, Jani S, ISBN: 0-944838-32-4, 1993, 172 pp, Medical Physics Publishing, Madison (WI).

Geometric Uncertainties in Radiotherapy, BIR, 2003, ISBN: 0905749537, The British Institute of Radiology.

Intraoperative Irradiation, Techniques and Results, Gunderson LL, Wilett CG, Harrison LB, Calvo FA, editors ISBN: 0-89603-523-9, 1999, 560 pp, Advanced Medical Publishing, Madison (WI).

Intraoperative Radiation Therapy, Ralph Dobelbower, Jr. and Mitsuyuki Abe, ISBN: 0849368464, 1989, 432 pp, CRC Press, Boca Raton (FL).

Introduction to Clinical Radiation Oncology, 3rd ed., Coia L, Moylan D, ISBN: 0-944838-70-7, Published: March 1998, 568 pp, Medical Physics Publishing, Madison (WI).

IPEM Report No. 68, A Guide to Commissioning & Quality Control of Treatment Planning Systems, Shaw J, editor, ISBN: 0904181839, 1996, IPEM, York (UK).

IPEM Report No. 81, Physics Aspects of Quality Control in Radiotherapy, Mayles WPM, Lake RA, McKenzie AL, Macaulay EM, Morgan HM, Powley SK, ISBN: 0904181928, 1998, IPEM, York (UK).

IPEM Report No. 83, Targeted Radiotherapy, Fleming JS, Perkins AC, ISBN: 0904181979, Published: 2000, 112 pp, IPEM, York (UK).

Linac and Gamma Knife Radiosurgery, Isabelle M. Germano, ISBN: 1879284707, 2000, 295 pp, Advanced Medical Publishing, Madison (WI).

Linear Accelerators for Radiation Therapy, D. Greene, ISBN: 0750304766, 1997, 288 pp, IOPP, Bristol (UK).

Monitor Unit Calculations for External Photon and Electron Beams, Gibbon JP, editor, ISBN: 1-883526-08-6, 2000, 152 pp, Advanced Medical Publishing, Madison (WI).

Physical Aspects of Brachytherapy, Godden TJ, ISBN: 0852745117, 1988, 304 pp, IOPP, Bristol (UK).

Physical Aspects of Stereotactic Radiosurgery, Phillips M H, ISBN: 0306445352, 1993, 286 pp, Plenum, New York.

Physics and Technology of Hyperthermia, Field SB, Franconi C, ISBN: 9024735092, 1999, 668 pp, Springer.

Physics of Electron Beam Therapy, Klevenhagen SC, ISBN: 0852747810, 1985, 214 pp, IOPP, Bristol (UK).

Physics of Radiotherapy X-Rays from Linear Accelerators, Metcalfe P et al., ISBN: 0-944838-76-6, 1997, 493 pp, Medical Physics Publishing, Madison (WI).

Practical Essentials of Intensity Modulated Radiation Therapy, Chao KSC, Smith Apisarnthanarax, and Gokhan Ozyigit, ISBN: 0-7817-5279-5, 2004, 324 pp, Advanced Medical Publishing, Madison (WI).

Practical Manual of Brachytherapy, Pierquin B, Marinello G, ISBN: 0-944838-73-1, 1997, 296 pp, Medical Physics Publishing, Madison (WI).

Primer on Theory and Operation of Linear Accelerators, 2nd ed., Karzmark CJ, Morton R, ISBN: 0-944838-66-9, 1998, 50 pp, Medical Physics Publishing, Madison (WI).

Principles and Practice of Brachytherapy, Nag S, editor, ISBN: 0879936541, 1997, 752 pp, Futura.

Principles and Practice of Brachytherapy Using Afterloading Systems, Joslin CA, Flynn A, Hall EJ, editors, ISBN: 0-340-74209-7, 2001, 464 pp, Edward Arnold, London.

Protocol and Procedures for Quality Assurance of Linear Accelerators, Constantinou, ISBN: 0-9638266-0-3, 1993, 92 pp, Medical Physics Publishing, Madison (WI).

Quality Assurance in Radiotherapy Physics, Starkschall G, ISBN: 0944838219, 1991, 387 pp, Medical Physics Publishing, Madison (WI).

Radiation Therapy Planning, Bentel GC, ISBN: 0070051151, 1995, 643 pp, McGraw-Hill, New York.

Radiotherapy In Practice – Brachytherapy, Hoskin PJ, Coyle C, ISBN: 0198529406, 2005, 224 pp, Oxford University Press, New York.

Study Guide for Radiation Oncology Physics Board Exams, Berman B, Thomadsen B, ISBN: 0-944838-94-4, 2000, 112 pp, Medical Physics Publishing, Madison (WI).

The Physics and Radiobiology of Fast Neutron Beams, Bewley DK, ISBN: 085274093x, 1989, 192 pp, IOPP, Bristol (UK).

The Physics of Conformal Radiotherapy: Advances in Technology, Webb S, ISBN: 0750303972, 1997, 382 pp, IOPP, Bristol (UK).

The Physics of Modern Brachytherapy for Oncology, Baltas D, Kreiger H, Zamboglou N, ISBN: 0750307080, 2005, 450 pp, IOPP, Bristol (UK).

The Physics of Three Dimensional Radiation Therapy: Conformal Radiotherapy, Radiosurgery and Treatment Planning, Webb S, ISBN: 075030247x, 1993, 373 pp, IOPP, Bristol (UK).

The Q Book – The Physics of Radiotherapy X-Rays: Problems and Solutions Metcalfe P et al., ISBN: 0-944838-86-3, 1998, 100 pp, Medical Physics Publishing, Madison (WI).

The Theory & Practice of Intensity Modulated Radiation Therapy, Sternick S, editor, ISBN: 1-883526-05-1, 1997, 256 pp, Advanced Medical Publishing, Madison (WI).

The Use of Computers in Radiation Therapy: Schlegel W, Bortfeld T, editors, ISBN: 3540671765, 2000, 604 pp, Springer, New York.

The Use of Plane Parallel Ionization Chambers in High Energy Electron and Photon Beams: An International Code of Practice for Dosimetry, IAEA, ISBN: 9201048963, 1997, 125 pp, IAEA.

Therapy Physics Review, Paliwal B, ISBN: 0-944838-67-7, 1996, 65 pp, Medical Physics Publishing, Madison (WI).

Three-Dimensional Radiation Treatment: Technological Innovations and Clinical Results, Kneschaurek P, Molls M, Feldmann HJ, ISBN: 3805569475, 2000, S. Karger.

Topics in Dosimetry & Treatment Planning for Neutron Capture Therapy, Zamenhof RG, Solares GR, Harling OK, editors, ISBN: 1-883526-02-7, 1994, 245 pp, Advanced Medical Publishing, Madison (WI).

Treatment Planning in Radiation Oncology, Khan F M, Potish R, editors, ISBN: 0-683-04607-1, 1997, 608 pp, Lippincott, New York.

DIAGNOSTIC RADIOLOGICAL PHYSICS

General

AAPM Monograph No. 3, The Physics of Medical Imaging: Recording System Measurements and Techniques (1979 Summer School), Haus AG, editor, ISBN: 0-88318-260-2, 624 pp, Medical Physics Publishing, Madison (WI).

AAPM Monograph No. 23, The Expanding Role of Medical Physics in Diagnostic Imaging, Frey GD, Sprawls P, editors, ISBN: 1-888340-09-6, 1997, 583 pp, Medical Physics Publishing, Madison (WI).

Christensen's Physics of Diagnostic Radiology, Curry T S III, Dowdey JE, Murry RC Jr, ISBN: 0812113101, 1990, 522 pp, Lippincott, New York.

IPEM Report 61, Physics in Diagnostic Radiology, Faulkner K, Cranley K, Starritt HC, Wankling PF, editors, ISBN: 090418160X, 1990, 150 pp, IPEM, York (UK).

Physics for Diagnostic Radiology, Dendy P P, Heaton B, ISBN: 0750305916, 1999, 446 pp, IOPP, Bristol (UK).

Practical Radiography, Robert Ward, ISBN: 0-944838-49-9, (1996 reprint), 112 pp, Medical Physics Publishing, Madison (WI).

The Physics of Diagnostic Imaging, Dowsett DJ, Johnston RE, Kenny PA, ISBN: 0412460602, 1998, 609 pp, Edward Arnold, London.

Topical

AAPM Monograph No. 4, Quality Assurance in Diagnostic Radiology, Waggener R, Wilson C, editors, ISBN: 0-883182-68-8, 1977, 190 pp, Medical Physics Publishing, Madison (WI).

AAPM Monograph No. 30, Specifications, Performance Evaluation and Quality Assurance of Radiographic and Fluoroscopic Systems in the Digital Era, Goldman L, Yester M, editors, ISBN: 1-930524-21-8, 2004, 300 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 4, Basic Quality Control in Diagnostic Radiology, Task Force On Quality Assurance Protocol; ISBN: 0-88318-251-3, 1977, 57 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 14, Performance Specifications and Acceptance Testing for X-Ray Generators and Automatic Exposure Control Devices, AAPM, ISBN: 0-88318-461-3, 1985, 96 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 15, Performance Evaluation and Quality Assurance in Digital Subtraction Angiography, Diagnostic X-Ray Imaging Committee/ DigitalRadiography/Fluorography Task Group; ISBN: 0-88318-482-6, 1985, 36 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 29, Equipment Requirements and Quality Control for Mammography, Diagnostic X-Ray Imaging Committee Task Group 7, ISBN: 0-88318-807-4, 1990, 72 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 31, Standardized Methods for Measuring Diagnostic X-Ray Exposures, Diagnostic X-Ray Imaging Committee Task Group 8, ISBN: 0-88318-874-0, 1990, 22 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 58, Managing the Use of Fluoroscopy in Medical Institutions, Radiation Protection Committee Task Group 6; ISBN: 1-888340-13-4, 1998, 42 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 60, Instrumentation Requirements of Diagnostic Radiological Physicists, Diagnostic X-Ray Imaging Committee Task Group 4; ISBN: 1-888340-15-0, 1998, 40 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 70, Cardiac Catheterization Equipment Performance, Diagnostic X-ray Imaging Committee, Task Group 17, ISBN: 1-888340-28-2, 2001, 71 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 74, Quality Control in Diagnostic Radiology, Task Group 12 ISBN: 1-888340-33-9, 2002, 77 pp, Medical Physics Publishing, Madison (WI).

Advances in Film Processing Systems Technology and Quality Control in Medical Imaging, Haus AG, ISBN: 1-930524-01-3, 2001, 245 pp., Medical Physics Publishing, Madison (WI).

Basics of Film Processing in Medical Imaging, Haus A, Jaskulski S, ISBN: 0-944838-78-2, 1997, 338 pp, Medical Physics Publishing, Madison (WI).

Digital Mammography Proceedings, Yaffe M, ISBN: 1-930524-00-5, 2001, 856 pp Medical Physics Publishing, Madison (WI).

Interventional Fluoroscopy: Physics, Technology, Safety, Balter S, ISBN: 0471390100, 2001, 284 pp, Wiley, New York.

IPEM Report No. 32, Measurement of the Performance Characteristics of Diagnostic X-Ray Systems Used in Medicine. Part 1: X-Ray Tubes and Generators, Cranley K ISBN: 090418174X, 1995, 28 pp, IPEM, York (UK).

IPEM Report No. 32, Measurement of the Performance Characteristics of Diagnostic X-Ray Systems Used in Medicine. Part II: X-Ray Image Intensifier Television Systems, Starritt H C, ISBN: 0904181758, 1996, 61 pp, IPEM, York (UK).

IPEM Report No. 32, Measurement of the Performance Characteristics of Diagnostic X-Ray Systems Used in Medicine. Part IV: X-Ray Intensifying Screens, Films, Processors and Automatic Exposure Control Systems, Holubinka M R, ISBN: 0904181774, 1996, 43 pp, IPEM, York (UK).

IPEM Report No. 32, Measurement of the Performance Characteristics of Diagnostic X-Ray Systems Used in Medicine. Part V: Conventional Tomographic Equipment, ISBN: 0904181782, 1996, 18 pp, IPEM, York (UK).

IPEM Report No. 32, Measurement of the Performance Characteristics of Diagnostic X-Ray Systems Used in Medicine. Part VI: X-Ray Image Intensifier Fluorography Systems, Robertson J, ISBN: 0904181790, 1995, 21 pp, IPEM, York (UK).

IPEM Report No. 59, The Commissioning & Routine Testing of Mammographic X-Ray Systems 2nd ed., Law J, Dance DR, Faulkner K, Fitzgerald MC, Ramsdale ML, Robinson A, ISBN: 0904181723, 1994, 119 pp, IPEM, York (UK).

IPEM Report No. 67 Quality Assurance in Dental Radiology, Starritt HC, Faulkner K, Wankling PF, Cranley K, Robertson J, Young K, ISBN: 0904181677, 1994, 25 pp, IPEM, York (UK).

IPEM Report No. 77, Recommended Standards for Routine Testing of Diagnostic X-Ray Imaging Systems, ISBN: 0904181871, 1997, 64 pp, IPEM, York (UK).

IPEM Report No. 78, Catalogue of Diagnostic X-Ray Spectra & Other Data Cranley K, Gilmore BJ, Fogarty GWA, Desponds L, ISBN: 090418188X, 1997, IPEM, York (UK).

IPEM Report No. 79, The Critical Examination of X-Ray Generating Equipment in Diagnostic Radiology, ISBN: 0904181898, 1998, 17 pp, IPEM, York (UK).

Mammography Quality Control: The Why & How Book, Calvin Myers, ISBN: 0-944838-83-9, 1997, 43 pp, Medical Physics Publishing, Madison (WI).

Medical Imaging and Radiation Protection for Medical Students and Clinical Staff, Martin CJ, Dendy PP, Corbett RH, 2003, The British Institute of Radiology.

Practical Digital Imaging and PACS (1999 AAPM Summer School), Seibert A et al., ISBN: 0-944838-92-8, 1999, 577 pp, Medical Physics Publishing, Madison (WI).

Principles of Radiographic Imaging: An Art and a Science, Carlton RR, Adler A, ISBN: 0766813002, 2000, 752 pp, Thomson Delmar Learning.

Radiation Exposure and Image Quality in X-Ray Diagnostic Radiology: Physical Principles and Clinical Applications, Aichinger H, Dierker J, Joite-Barfuß S, Säbel M, ISBN: 3540442871, 2003, 212 pp, Springer, New York.

Radiology Review: Radiologic Physics, Nickoloff EL, Naveed Ahmad, ISBN: 1416022600, 2005, 272 pp, Saunders, Philadelphia.

Screen Film Mammography: Imaging Considerations and Medical Physics Responsibilities, Gary Barnes and G. Donald Frey, ISBN: 0-944838-12-X, 1991, 127 pp, Medical Physics Publishing, Madison (WI).

IMAGING

General

3D Imaging in Medicine, 2nd ed., Udupa JK, Herman GT, ISBN: 084933179X, 1999, 384 pp, CRC Press, Boca Raton (FL).

Essentials of Diagnostic Imaging, Guebert G M, Pirtle OL, Yochum TR, ISBN: 0801674557, 1995, 252 pp, Mosby, Philadelphia.

Foundations of Image Science, Barrett HH, Myers K, ISBN: 0471153001, 2003, 1100 pp, Wiley, New York.

Fundamentals of Medical Imaging, Paul Suetens, ISBN: 0521803624, 2002, 294 pp, Cambridge University Press, New York.

Handbook of Medical Imaging, Volume 1: Physics and Psychophysics, Beutel J, Kundel HL, Van Metter RL, ISBN: 0819436216, 2000, 968 pp, SPIE.

Introduction to Biomedical Imaging, Webb AG, ISBN: 0471237663, 2002, 264 pp, Wiley, New York.

Introduction To The Principles of Medical Imaging, Guy C, ISBN: 1860945023, 2005, 400 pp, Imperial College Press, London.

Medical Imaging 2004: Physics of Medical Imaging (SPIE Proceedings), Yaffe MJ, ISBN: 0819452815, 2004, SPIE.

Medical Imaging Physics, 4th ed., Hendee WR, Ritenour ER, ISBN: 0-471-38226-4, 2002, 536 pp, Wiley, New York.

Physics for Medical Imaging, Farr RF, Allisy-Roberts PJ, ISBN: 0702017701, 1997, 276 pp, Bailliere Tindall.

Physics of Medical Imaging, Dobbins JT, Boone JM, ISBN: 0819427810, 1998, 842 pp, SPIE.

Principles of Imaging Science and Protection, Thompson MA, Hattaway MP, Hall JD, ISBN: 0721634281, 1994, 522 pp, Saunders, Philadelphia.

Principles of Medical Imaging, Kirk Shung K, Smith MB, Tsui BMW, ISBN: 0126409706, 1992, 289 pp, Academic Press.

The Essential Physics of Medical Imaging, Hardbound, 2nd ed., Bushberg JT, Seibert JA, Leidholdt EM Jr, Boone JM, ISBN: 0-683-30118-7, 2001, 965 pp, Lippincott.

The Physics of Diagnostic Imaging, 2nd ed., Dowsett DJ, Kenny PA, Johnston RE, ISBN: 0412460602, 2005, Hodder Headline (Arnold).

The Physics of Medical Imaging, Webb S, ISBN: 0852743491, 1988, 633 pp, IOPP, Bristol (UK).

Topical

AAPM Monograph No. 11, Electronic Imaging in Medicine, Fullerton GD, Hendee W, Lasher J, Properzio W, Riederer S, editors, ISBN: 0-88318-454-0, 1983, 484 pp, Medical Physics Publishing, Madison (WI).

AAPM Monograph No. 12, Recent Developments in Digital Imaging (1984 Summer School), Doi K, Lanzl L, Lin P-J P, editors, ISBN: 0-88318-463-X, 1984, 576 pp, Medical Physics Publishing, Madison (WI).

AAPM Monograph No. 25, Practical Digital Imaging and PACS (1999 AAPM Summer School), Seibert A et al., ISBN: 0-944838-92-8, 1999, 577 pp, Medical Physics Publishing, Madison (WI).

Electrical Impedance Tomography: Methods, History and Applications, Holder DS, ISBN: 0750309520, 2005, 456 pp, IOPP, Bristol (UK).

Mathematics of Medical Imaging, Epstein CL, ISBN: 0130675482, 2003, 768 pp, Prentice Hall, New York.

Medical Imaging Signals and Systems, Prince JL, Links J, ISBN: 0130653535, 2005, 550 pp, Prentice Hall, New York.

Wavelet Analysis with Applications to Image Processing, Lakshman Prasad and S. Sitharama Iyengar, ISBN: 0849331692, 1997, 304 pp, CRC Press, Boca Raton (FL).

COMPUTERIZED TOMOGRAPHY

General

AAPM Monograph No. 6, Medical Physics of CT and Ultrasound: Tissue Imaging and Characterization (1980 Summer School), Fullerton GD, Zagzebski J, editors, ISBN: 1-888340-08-8, 1980, 717 pp, Medical Physics Publishing, Madison (WI).

Computed Tomography: Fundamentals, System Technology, Image Quality, Applications, Kalender WA, ISBN: 3-8957-8081-2, 2000, 220 pp, Advanced Medical Publishing, Madison (WI).

CT Physics: The Basics, Villafana T, ISBN: 0683307118, 2002, 250 pp, Lippincott.

Topical

AAPM Report No. 1, Phantoms for Performance Evaluation and Quality Assurance of CT Scanners, AAPM, ISBN: 1-888340-04-5, 1977, 23 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 39, Specification and Acceptance Testing of Computed Tomography Scanners, Diagnostic X-Ray Imaging Committee Task Group 2; ISBN: 1-56396-230-6, 1993, 95 pp, Medical Physics Publishing, Madison (WI).

IPEM Report No. 32, Measurement of the Performance Characteristics of Diagnostic X-Ray Systems Used in Medicine. Part III: Computed Tomography X-Ray Scanners, ISBN: 0904181766, 2003, 94 pp, IPEM, York (UK).

NUCLEAR MEDICINE

General

AAPM Monograph No. 10, Physics of Nuclear Medicine: Recent Advances (1983 Summer School), Rao D, Chandra R, Graham M, editors, ISBN: 0-88318-440-0, 1983, 560 pp, Medical Physics Publishing, Madison (WI).

Diagnostic Nuclear Medicine: A Physics Perspective, Hamilton DI, ISBN: 3540006907, 2004, 465 pp, Springer, New York.

Essentials of Nuclear Medicine Physics, Powsner RA, Powsner ER, ISBN: 0632043148, 1998, 199 pp, Blackwell, Cambridge (MA).

Handbook of Nuclear Medicine, Madsen M, Ponto J, ISBN: 0-944838-14-6, 1992, 114 pp, Medical Physics Publishing, Madison (WI).

Introductory Physics of Nuclear Medicine, Ramesh Chandra, ISBN: 0812114426, 1992, 221 pp, Lea & Febiger, Philadelphia (PA).

Nuclear Medicine and PET: Technology and Techniques, Christian PE, Bernier D, Langan JK, ISBN: 0323019641, 2003, 640 pp, Mosby.

Nuclear Medicine Physics: The Basics, Ramesh Chandra, ISBN: 068330092X, 1998, 182 pp, Lippincott, New York.

Physics in Nuclear Medicine, Cherry SR, Sorenson J, Phelps M, ISBN: 072168341X, 2003, 523 pp, Saunders, Philadelphia.

Practical Nuclear Medicine, 3rd ed., Sharp PF, Gemmell HG, Murray AD, ISBN: 185233875X, 2005, 352 pp, Springer, New York.

Principles and Practice of Nuclear Medicine, Early PJ, Bruce D,. Sodee MD, ISBN: 0801625777, 1995, 877 pp, Mosby.

Topical

AAPM Manual No. 1: Nuclear Medicine Instrumentation Laboratory Exercises for Radiology Residency Training, Van Tuinen R J, Grossman LW, Kereiakes JG, editors, ISBN: 0-88318-0001, 1994, 81 pp, Medical Physics Publishing, Madison (WI).

AAPM Monograph No. 1, Biophysical Aspects of Medical Use of Technetium-99m, Kereiakes JG, Corey KR, editors, ISBN: 1-888340-05-3, 1976, 126 pp, Medical Physics Publishing, Madison (WI).

AAPM Monograph No. 18, Expanding the Role of Medical Physics in Nuclear Medicine (1989 Summer School), Frey GD, Yester MV, editors, ISBN: 0-883189-15-1, 1989, 368 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 6, Scintillation Camera Acceptance Testing & Performance Evaluation, Nuclear Medicine Committee; ISBN: 0-88318-275-0, 1980, 23 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 9, Computer-Aided Scintillation Camera Acceptance Testing, Nuclear Medicine Committee Task Group; ISBN: 0-88318-407-9, 1981, 40 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 22, Rotating Scintillation Camera SPECT Acceptance Testing and Quality Control, Nuclear Medicine Committee Task Group; ISBN: 0-88318-549-0, 1987, 26 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 52, Quantitation of SPECT Performance (Reprinted from Medical Physics, Vol. 2, Issue 4), Nuclear Medicine Committee Task Group 4; ISBN: 1-56396-485-6, 1995, 10 pp, Medical Physics Publishing, Madison (WI).

IPEM Report No. 65, Quality Standards in Nuclear Medicine, Edited by Hart GC and Smith AH, ISBN: 0904181642, 1992, 123 pp, IPEM, York (UK).

IPEM Report No. 66, Quality Control of Gamma Cameras & Associated Computer Systems, Hannan J, editor, ISBN: 0904181650, 1992, 62 pp, IPEM, York (UK).

IPEM Report No. 85, Radioactive Sample Counting – Principles and Practice, Driver I, editor, ISBN: 0904181995, 2002, 63 pp, IPEM, York (UK).

IPEM Report No. 86, Quality Control of Gamma Camera Systems, Bolster A, ISBN: 1903613132, 2003, 130 pp, IPEM, York (UK).

IPEM Report No. 87, Basics of Gamma Camera Positron Emission Tomography Hillel P, editor, ISBN: 1903613183, 2004, 73 pp, IPEM, York (UK).

Positron Emission Tomography: Basic Sciences, Bailey DL, Townsend DW, Valk PE, Maisey MN, ISBN: 1852337982, 2005, 382 pp, Springer, New York.

Principles and Practice of Positron Emission Tomography, Wahl RL, ISBN: 0781729041, 2002, 442 pp, Lippincott, New York.

Therapeutic Applications of Monte Carlo Calculations in Nuclear Medicine, Habib Zaidi, ISBN: 0750308168, 2003, 363 pp, IOPP, Bristol (UK).

MAGNETIC RESONANCE IMAGING AND SPECTROSCOPY

General

AAPM Monograph No. 14, NMR in Medicine: The Instrumentation and Clinical Applications (1985 Summer School), Thomas SR, Dixon RL, editors, ISBN: 0-88318-497-4, 1985 595 pp, Medical Physics Publishing, Madison (WI).

AAPM Monograph No. 21, The Physics of MRI, Michael Bronskill, Sprawls P, editor, ISBN: 1-563962-05-5, 1992, 784 pp, Medical Physics Publishing, Madison (WI).

Magnetic Resonance Imaging: Physical Principles and Sequence Design, Haacke EM et al., ISBN: 0471351288, 1999, 914 pp, John Wiley & Sons, Inc., New York.

Magnetic Resonance Imaging: Principles, Methods, and Techniques, Sprawls P, ISBN: 0-944838-97-9, 2000, 200 pp, Medical Physics Publishing, Madison (WI).

Magnetic Resonance Imaging: Theory and Practice, Vlaardingerbroek MT, Den Boer JA, ISBN: 3540600809, 1996, 347 pp, Springer, New York.

Magnetic Resonance in Medicine, 4th ed., Peter Rinck, ISBN: 0632059869, 2001, 245 pp, Blackwell.

Non-Mathematical Approach to Basic MRI, Smith H, Ranallo F, ISBN: 0-944838-02-2, Published: 1989, 203 pp, Medical Physics Publishing, Madison (WI).

Topical

AAPM Report No. 20, Site Planning for Magnetic Resonance Imaging Systems, Nuclear Magnetic Resonance Committee Task Group 2; ISBN: 0-88318-530-X, 1986, 60 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 28, Quality Assurance Methods and Phantoms for Magnetic Resonance Imaging (Reprinted from Medical Physics, Vol. 17, Issue 2), Nuclear Magnetic Resonance Committee Task Group 1, ISBN: 0-88318-800-7, 1990, 9 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 34, Acceptance Testing of Magnetic Resonance Imaging Systems (Reprinted from Medical Physics, Vol. 19, Issue 1), Nuclear Magnetic Resonance Committee Task Group 6; ISBN: 1-56396-028-1, 1992, 13 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 77, Practical Aspects of Functional MRI, Nuclear Medicine Committee Task Group 8, ISBN: 1-888340-37-1, 2002, 22 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 78, Proton Magnetic Resonance Spectroscopy in the Brain, Magnetic Resonance Task Group 9, ISBN: 1-888340-38-X, 2002, 21 pp, Medical Physics Publishing, Madison (WI).

Handbook of MRI Pulse Sequences, Matt Bernstein, Kevin King, Xiaohong Joe Zhou, ISBN: 0120928612, 2004, 1040 pp, Academic Press.

Introduction to Functional Magnetic Resonance Imaging: Principles and Techniques, Buxton RB, ISBN: 0521581133, 2001, 536 pp, Cambridge University Press, New York.

IPEM Report No. 80, Quality Control in Magnetic Resonance Imaging, RA Lerski, J De Wilde, D Boyce and J Ridgeway, ISBN: 0904181901, 1999, 50 pp, IPEM, York (UK).

Principles of Magnetic Resonance Imaging: A Signal Processing Perspective, Liang Z-P, Lauterbur PC, ISBN: 0780347234, 1999, 416 pp, Wiley, New York.

ULTRASOUND PHYSICS

General

AAPM Monograph No. 6, Medical Physics of CT and Ultrasound: Tissue Imaging and Characterization (1980 Summer School), Fullerton GD, Zagzebski J, editors, ISBN: 1-888340-08-8, 1980, 717 pp, Medical Physics Publishing, Madison (WI).

Advances in Ultrasound Techniques and Instrumentation, Wells PNT, ISBN: 0443088535, 1993, 192 pp, Saunders, Philadelphia.

Clinical Ultrasound Physics: A Workbook for Physicists, Residents, and Students, Kofler JMJr, et al., ISBN: 1-930524-06-4a, 2001, 85 pp, Medical Physics Publishing, Madison (WI).

Diagnostic Ultrasound: Physics and Equipment, Hoskins P, Thrush A, Martin K, Whittingam T, Hoskins PR, Thrush A, Whittingham T, ISBN: 1841100420, 2002, 208 pp, Cambridge University Press, New York.

Essentials of Ultrasound Physics, Zagzebski J A, ISBN: 0815198523, 1996, 220 pp, Mosby.

Physical Principles of Medical Ultrasonics, Hill CR, Bamber JC, ter Haar GR, ISBN: 0471970026, 2002, 528 pp, Wiley, New York.

Physics and Instrumentation of Diagnostic Medical Ultrasound, Fish P, ISBN: 0471958956, 2005, 250 pp, Wiley, New York.

Principles and Applications of Ultrasound, Langton CM, ISBN: 0750308052, 2005, 252 pp, IOPP, Bristol (UK).

The Physics of Clinical MR Taught Through Images, Runge VM, Nitz WR, Schmeets SH, Faulkner WH, Desai NK, ISBN: 1588903222, 2004, 221 pp, Thieme Med. Pub.

Ultrasound in Medicine, Duck FA, Baker AC, Starritt HC, editors, ISBN: 0750305932, 1998, 314 pp, IOPP, Bristol (UK).

Ultrasound Physics Mock Exam, Owen CA, Zagzebski JA, ISBN: 0941022633, 2004, Davies, Inc.

Topical

AAPM Report No. 8, Pulse Echo Ultrasound Imaging Systems: Performance Tests and Criteria, General Medical Physics Committee Ultrasound Task Group; ISBN: 0-88318-283-1, 1980, 73 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 65, Real-Time B-Mode Ultrasound Quality Control Test Procedures (Reprinted from Medical Physics, Vol. 25, Issue 8), Ultrasound Task Group 1; ISBN: 1-888340-22-3, 1998, 22 pp, Medical Physics Publishing, Madison (WI).

Basic Doppler Physics, Smith H, Zagzebski J, ISBN: 0-944838-15-4, 1991, 136 pp, Medical Physics Publishing, Madison (WI).

Doppler Ultrasound: Physics, Instrumental, and Clinical Applications, 2nd ed., Evans DH, McDicken WN, ISBN: 0471970018, 2000, 456 pp, Wiley, New York.

IPEM Report No. 70, Testing of Doppler Ultrasound Equipment, Hoskins PR, Sherriff SB, Evans JA, ISBN: 0904181715, 1994, 136 pp, IPEM, York (UK).

IPEM Report No. 71, Routine Quality Assurance of Ultrasound Imaging Systems ISBN: 0904181820, 1995, 66 pp, IPEM, York (U.K.).

IPEM Report No. 84, Guidelines for the Testing and Calibration of Physiotherapy Ultrasound Machines, Pye S, Zequiri B, ISBN: 0904181987, 2001, 67 pp, IPEM, York (UK).

Safety of Diagnostic Ultrasound, Barnett SB, Kossoff G, editors, ISBN: 1850706468, 1997, 147 pp, Taylor & Francis.

The Safe Use of Ultrasound in Medical Diagnosis, ter Haar G, Duck FA, ISBN 0-905749-42-1, 2000, 120 pp, British Medical Ultrasound Society and British Institute of Radiology, London (UK).

Three-Dimensional Ultrasound, Downey B, Pretorius DH, Fenster A, ISBN: 0-7817-1997-6, 1999, 272 pp, Lippincott Williams & Wilkins, Philadelphia.

LIGHT AND LASERS

General

AAPM Report No. 3, Optical Radiations in Medicine: A Survey of Uses, Measurement and Sources, AAPM, ISBN: 1-888340-06-1, 1977, 28 pp, Medical Physics Publishing, Madison (WI).

An Introduction to Biomedical Optics, Splinter R, ISBN: 0750309385, 2005, 350 pp, IOPP, Bristol (UK).

Applied Laser Medicine, Breuer H, Krasner N, Okunata T, Sliney D, Berlien H-P, Müller GJ, ISBN: 354067005X, 2004, 740 pp, Springer, New York.

Laser-Tissue Interactions: Fundamentals and Applications, Niemz MH, ISBN: 3540405534, 2003, 305 pp, Springer, New York.

Light, Visible and Invisible, and Its Medical Applications, Newing A, ISBN: 1860941648, 1999, 212 pp, World Scientific, River Edge (NJ).

Topical

AAPM Report No. 57, Recommended Nomenclature for Physical Quantities in Medical Applications of Light, General Medical Physics Committee Task Group 2; ISBN: 1-888340-02-9, 1996, 6 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 73, Medical Lasers: Quality Control, Safety, Standards, and Regulations, Task Group 6, ISBN: 1-888340-31-2, 2001, 68 pp, Medical Physics Publishing, Madison (WI).

IPEM Report No. 76, Ultraviolet and Blue-Light Phototherapy-Principles, Sources, Dosimetry and Safety, Diffey B, Hart G, ISBN: 0904181863, Published: 1997, 56 pp, IPEM, York (U.K.).

Laser Safety, Henderson R, ISBN: 0750308591, 2003, 480 pp, IOPP, Bristol (U.K.).

Laser Systems for Photobiology and Photomedicine, Chester AN, Martellucci S, Scheggi AM, ISBN: 0306438860, 1991, 311 pp, Plenum, New York.

Ultraviolet Radiation in Medicine, Diffey BL, ISBN: 0852745354, 1982, 172 pp, IOPP, Bristol (U.K.).

RADIATION PROTECTION

General

Atoms, Radiation, and Radiation Protection, 2nd ed., Turner JE, ISBN: 0-471-59581-0, 1995, 576 pp, Wiley, New York.

Basic Health Physics: Problems and Solutions, Bevelacqua J, ISBN: 0471297119, 1999, 559 pp, Wiley, New York.

Basic Radiation Protection Technology (2nd edition), Gollnick DA, ISBN: 0916339033, 1988, Pacific Radiation Corp.

Contemporary Health Physics: Problems and Solutions, Bevelacqua J, ISBN: 0471018015, 1995, 456 pp, Wiley, New York.

CRC Handbook of Management of Radiation Protection Programs, 2nd ed., Miller KL, ISBN: 0849337704, 1992, 496 pp, CRC Press, Boca Raton (FL).

Exposure Criteria for Medical Diagnostic Ultrasound: 2. Criteria Based on All Known Mechanisms (NCRP Report, No. 140), ISBN: 0929600738, 2003, NCRP.

Handbook of Health Physics and Radiological Health, Shleien B, Slaback LA Jr, Birky B, ISBN: 0683183346, 1997, 700 pp, Lippincott.

Introduction to Health Physics, 3rd ed., Cember H, ISBN: 00-71054618, 1996, 731 pp, McGraw-Hill, New York.

Medical Effects of Ionizing Radiation, 2nd ed., Mettler FA Jr, Upton AC, ISBN: 0721666469, 1995, 440 pp, Elsevier.

Physics for Radiation Protection, Martin JE, ISBN: 0-471-35373-6, 2000, 713 pp, Wiley, New York.

Practical Radiation Protection and Applied Radiobiology, 2nd ed., Dowd SB, Tilson ER, editors, ISBN: 0721675239, 1999, 368 pp, Saunders, New York.

Practical Radiation Protection in Healthcare, Martin CJ, Sutton DG, editors, ISBN: 0192630822, 2002, 440 pp, Oxford University Press, New York.

Radiation Protection, Seeram E, Travis E, ISBN: 0-397-55032-4, 1996, 320 pp, Lippincott.

Radiation Protection, Kathren RL, ISBN: 0852745540, 1985, 212 pp, IOPP, Bristol (UK).

Radiation Protection, Hallenbeck WH, ISBN: 0873719964, 1994, 288 pp, CRC Press, Boca Raton (FL).

Radiation Protection: A Guide for Scientists and Physicians, Shapiro J, ISBN: 0674745868, 1990, 494 pp, Harvard University Press, Cambridge (MA).

Radiofrequency Radiation Standards: Biological Effects, Dosimetry, Epidemiology, and Public Health Policy, Klauenberg BJ, Grandolfo M, Erwin DN, ISBN: 0306449196, 1995, 476 pp, Kluwer, Norwell (MA).

Topical

AAPM Proceedings No. 4, Radiotherapy Safety, Thomadsen B, editor, ISBN: 0-88318-443-5, Published: 1984, 169 pp, Medical Physics Publishing. Madison (WI).

AAPM Report No. 19, Neutron Measurements Around High Energy X-Ray Radiotherapy Machines, Radiation Therapy Committee Task Group 27, ISBN: 0-88318-518-0, 1986, 34 pp, Medical Physics Publishing. Madison (WI).

AAPM Report No. 25, Protocols for the Radiation Safety Surveys of Diagnostic Radiological Equipment, Diagnostic X-Ray Imaging Committee Task Group 1; ISBN: 0-88318-574-1, 1988, 55 pp, Medical Physics Publishing. Madison (WI).

AAPM Report No. 35, Recommendations on Performance Characteristics of Diagnostic Exposure Meters (Reprinted from Medical Physics, Vol. 19, Issue 1), Diagnostic X-Ray Imaging Committee Task Group 6; ISBN: 1-56396-029-X, 1992, 11 pp, Medical Physics Publishing. Madison (WI).

Exposure of the Pregnant Patient to Diagnostic Radiations, Wagner L, et al., ISBN: 0-944838-72-3, 1997, 259 pp, Medical Physics Publishing, Madison (WI).

How Clean is Clean, How Safe is Safe? Eisenbud M, ISBN: 0-944838-33-2, 1993, 63 pp, Medical Physics Publishing. Madison (WI).

IPEM Report No. 50, Chernobyl: Response of Medical Physics Departments in the UK, Haywood JK, editor, ISBN: 0904181456, 1986, 99 pp, IPEM, York (UK).

IPEM Report No. 63, Radiation Protection in Nuclear Medicine & Pathology Goldstone KE, Jackson PC, Myers MJ, Simpson A E, editors, ISBN: 0904181626, 1991, 190 pp, IPEM, York (UK).

IPEM Report No. 69, Recommendations for the Presentation of Type Test Data for Radiation Protection Instruments in Hospitals, The Radiation Protection Instrument Calibration Working Party of the IPEM, ISBN: 0904181707, 1994, 12 pp, IPEM, York (UK).

IPEM Report No. 72, Safety in Diagnostic Radiology, ISBN: 904181812, 1995, 119 pp, IPEM, York (UK).

IPEM Report No. 75, The Design of Radiotherapy Treatment Room Facilities, Stedeford B, Morgan HM, Mayles WPM, ISBN: 1903613855, 1997, 161 pp, IPEM, York (UK).

IPEM Report No. 82, Cost-Effective Methods of Patient Dose Reduction in Diagnostic Radiology, ISBN: 0904181944, 2001, 50 pp, IPEM, York (UK).

IPEM Report No. 88, Guidance on the Establishment and Use of Diagnostic Reference Levels for Medical X-Ray Examinations, ISBN: 1903613205, 2004, 44 pp, IPEM, York (UK).

Management and Administration of Radiation Safety Programs (HPS 1998 Summer School), Charles Roessler, ISBN: 0-944838-01-4, 1998, 603 pp, Medical Physics Publishing, Madison (WI).

Public Protection from Nuclear, Chemical, and Biological Terrorism, Brodsky A, Johnson RH, Goans RE, editors, ISBN: 1-930524-23-4, Published: July 2004, 872 pp, Medical Physics Publishing, Madison (WI).

Radiation Injuries, Gooden D, ISBN: 33333, 1991, 239 pp, Medical Physics Publishing, Madison (WI).

Radiation Protection Dosimetry: A Radical Reappraisal, Simmons J, Watt D, ISBN: 0-944838-87-1, 1999, 160 pp, Medical Physics Publishing, Madison (WI).

Radiation Protection in Medical Radiography, Statkiewicz Sherer MA, Visconti PJ, Ritenour RE, ISBN: 0323014526, 2002, 336 pp, Mosby.

Radiation Safety and ALARA Considerations for the 21st Century, HPS Midyear Symposium, 2001, ISBN: 1-930524-02-1, 2001, 280 pp, Medical Physics Publishing. Madison (WI).

Shielding Techniques for Radiation Oncology Facilities, 2nd ed., McGinley PH, ISBN: 1-930524-07-2, 2002, 184 pp, Medical Physics Publishing. Madison (WI).

Subject Dose in Radiological Imaging, Ng K-H, Bradley DA, Warren-Forward HM, ISBN: 0-444-82989-x, 1998, Elsevier.

The Invisible Passenger, Radiation Risks for People Who Fly, Barish RJ, ISBN: 188352606X, 1999, 119 pp, Advanced Medical Publishing, Madison (WI).

University Health Physics, Belanger R, Papin PJ, editors, ISBN: 1-930524-15-3, 2003, 408 pp, Medical Physics Publishing. Madison (WI).

RADIATION MEASUREMENTS

General

Applications of New Technology: External Dosimetry, Higginbotham JF, ISBN: 0-944838-68-5, 1996, 464 pp, Medical Physics Publishing, Madison (WI).

Fundamentals of Radiation Dosimetry, Green S, ISBN: 075030913x, 2005, 275 pp, IOPP, Bristol (UK).

Medical Radiation Detectors: Fundamental and Applied Aspects, Kember NF, ISBN: 0750303190, 1994, 236 pp, IOPP, Bristol (UK).

Radiation Detection and Measurement, 3rd ed., Knoll GF, ISBN: 0-471-07338-5, 1999, 816 pp, Wiley, New York.

Radiation Dosimetry: Physical and Biological Aspects, Orton CG, ISBN: 0306420562, 1986, 344 pp, Plenum, New York.

Radiation Instruments, Cember H, ISBN: 1-930524-03-X, 2001, 472 pp, Medical Physics Publishing, Madison (WI).

Topical

A Procedural Guide to Film Dosimetry, Yeo IJ, Kim JO, ISBN: 1-930524-19-6, 2004, 65pp, Medical Physics Publishing, Madison (WI).

AAPM Proceedings No. 11, Kilovoltage X-Ray Dosimetry for Radiotherapy and Radiobiology, Ma C-M, Seuntjens J, editors, ISBN: 1-888340-16-9, 1998, 220 pp, Medical Physics Publishing, Madison (WI).

AAPM Proceedings No. 13, Recent Developments in Accurate Radiation Dosimetry, Seuntjens JP, Mobit PN, editors, ISBN: 1-930524-12-9, 2002, 353 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 12, Evaluation of Radiation Exposure Levels in Cine Cardiac Catheterization Laboratories, Diagnostic Radiology Committee Cine Task Force; ISBN: 0-88318-439-7, 1984, 28 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 63, Radiochromic Film Dosimetry (Reprinted from Medical Physics, Vol. 25, Issue 11), Radiation Therapy Committee Task Group 55; ISBN: 1-888340-20-7, 1998, 23 pp, Medical Physics Publishing, Madison (WI).

Instrumentation, Measurements and Electronic Dosimetry, (HPS Midyear 2000) ISBN: 0-944838-93-6, 2000, 271 pp, Medical Physics Publishing, Madison (WI).

Internal Radiation Dosimetry (1994 HPS Summer School), Raabe O, ISBN: 0-944838-47-2, 1994, 667 pp, Medical Physics Publishing, Madison (WI).

Microdosimetry and Its Applications, Rossi HH, Zaider M, ISBN: 3540585419, 1996, 321 pp, Springer, New York.

Practical Applications of Internal Dosimetry, Bolch WE, editor, ISBN: 1-930524-09-9, 2002, 480 pp, Medical Physics Publishing, Madison (WI).

RADIATION BIOLOGY

General

An Introduction to Radiobiology, Nias AHW, ISBN: 0471975907, 1998, 400 pp, Wiley, New York.

Basic Clinical Radiobiology, Steel GG, ISBN: 0340807830, 2002, 266 pp, Edward Arnold, London.

Biological Risks of Medical Irradiations, Fullerton G, ISBN: 0883182793, 1987, 335 pp, American Institute of Physics, New York.

Effects of Atomic Radiation, Schull WJ, ISBN: 0471125245, 1995, 97 pp, Wiley, New York.

Handbook of Radiobiology, Prasad KN, ISBN: 0849325013, 1995, 352 pp, CRC Press, Boca Raton (FL).

Primer of Medical Radiobiology, Travis E, ISBN: 0815188374, 1989, 302 pp, Mosby, Philadelphia.

Radiobiology for the Radiologist, 5th ed., Hall E J, ISBN: 0-7817-2649-2, 2000, 608 pp, Lippincott.

Topical

A Compilation of Radiobiology Practice Examinations for Residents in Diagnostic Radiology and Radiation Oncology, Chapman JD, Shahabi S, Chapman BA, ISBN: 1 883526 09 4, 2000, 163 pp, Advanced Medical Publishing, Madison (WI).

AAPM Proceedings No. 7, Prediction of Response in Radiation Therapy: Analytical Models and Modelling, Paliwal B et al., ISBN: 0-883186-24-1, 1989, 757 pp, Medical Physics Publishing, Madison (WI).

AAPM Proceedings No. 9, Prediction of Response in Radiation Therapy: Radiosensitivity, Repopulation, Paliwal BR, Herbert D, Fowler JF, Kinsella TJ, editors, ISBN: 1-56396-271-3, 1993, 383 pp, Medical Physics Publishing, Madison (WI).

AAPM Proceedings No. 10, Volume & Kinetics in Tumor Control & Normal Tissue Complications: 5th International Conference on Dose, Time, and Fractionation in Radiation Oncology, Paliwal BR, Herbert D, editors, ISBN: 1-888340-11-8, 1998, 483 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 18, A Primer on Low-Level Ionizing Radiation and Its Biological Effects, Biological Effects Committee; ISBN: 0-88318-514-1, 1986, 103 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 43, Quality Assessment and Improvement of Dose Response Models: Some Effects of Study Weaknesses on Study Findings, Biological Effects Committee Task Group 1; ISBN: 0-944838-45-6, 1993, 351 pp, Medical Physics Publishing, Madison (WI).

Applied Radiobiology and Bioeffect Planning, Wigg D, ISBN: 1-930524-05-6, 2001, 486pp, Medical Physics Publishing, Madison (WI).

Biological Models and Their Applications in Radiation Oncology, Olsen KJ, ISBN: 10883526-03-5, 1994, 61 pp, Advanced Medical Publishing, Madison (WI).

BJR Supplement 26: Chronic Irradiation: Tolerance and Failure in Complex Biological Systems, 2002, The British Institute of Radiology.

Health Effects of Exposure to Low-level Ionizing Radiation, Hendee WR, Edwards FM, editors, ISBN: 0750303492, 1996, 640 pp, IOPP, Bristol (UK).

Health Effects of Exposure to Low Levels of Ionizing Radiation: BEIR V, National Research Council, ISBN: 0309039959, 1989, 436 pp, National Academies Press.

Physics and Radiobiology of Nuclear Medicine, Gopal Saha, 2nd ed., ISBN: 0387950214, 2003, 253 pp, Springer, New York.

Prediction of Tumor Treatment Response, Chapman DJ, Peters LJ, Withers RH, ISBN: 0080346898, 1989, 336 pp, Elsevier.

Radiation Oncology – Radiobiological and Physiological Perspectives, Awwad H K, ISBN: 0792307836, 1990, 688 pp, Kluwer, Norwell (MA).

The Radiation Biology of the Vascular Endothelium, Rubin DB, ISBN: 0849348404, 1997, 272 pp, CRC Press, Boca Raton (FL).

RADIOLOGICAL PHYSICS FOR RADIOLOGICAL TECHNOLOGISTS

General

Introduction to Radiologic Sciences and Patient Care, 3rd ed., Adler AM, Carlton RR, ISBN 0721697828, 2003, 520 pp, Saunders, Philadelphia.

Radiologic Physics and Radiographic Imaging, Bushong SC, ISBN: 0323032648, 2004, Mosby, Philadelphia.

Radiologic Science for Technologists: Physics, Biology and Protection, Bushong SC, ISBN: 0323025552, 2004, 704 pp, Mosby.

The Fundamentals of X-Ray and Radium Physics, Joseph Selman, ISBN: 0398058709, 1994, 637 pp, Charles C. Thomas.

Topical

Magnetic Resonance Imaging: Physical and Biological Principles, Bushong SC, ISBN: 0323014852, 2003, 528 pp, Mosby, Philadelphia.

MRI for Technologists, Woodward P, ISBN: 0071353186, 2000, 408 pp, McGraw-Hill, New York.

Principles and Practice of Radiation Therapy, Washington CM, Leaver D, ISBN: 0323017487, 2004, 992 pp, Elsevier, New York.

Rad Tech's Guide to Mammography: Physics, Instrumentation, and Quality Control, Jacobson DR, Seeram E, ISBN: 0632044993, 2001, 120 pp, Blackwell.

Radiation Protection: Essentials of Medical Imaging Series, Stewart C. Bushong, ISBN: 0070120137, 1998, 288 pp, McGraw-Hill, New York.

The Basic Physics of Radiation Therapy, Selman J, ISBN: 0398056854, 1990, 749 pp, Charles C. Thomas.

The Fundamentals of Imaging Physics and Radiobiology for the Radiologic Technologist, Selman J, ISBN: 0398069875, 2000, 484 pp, Charles C. Thomas.

MATHEMATICS AND STATISTICS

AAPM Monograph No. 13, Multiple Regression Analysis: Applications in the Health Sciences, Herbert DE, Meyers R H, editors, ISBN: 0-88318-490-7, 1984, 598 pp, Medical Physics Publishing, Madison (WI).

Chaos and the Changing Nature of Science and Medicine: An Introduction, Herbert D, editor, ISBN: 1-56396-442-2, 1995, American Institute of Physics, New York.

Mathematical and Computer Modeling of Physiological Systems, Rideout V, ISBN: 0-13-563354-0, 1991, 155 pp, Prentice Hall, New York.

COMPUTERS

General

AAPM Monograph No. 17, Computers in Medical Physics (1988 Summer School), Benedetto A R, Huang HK, Ragan DP, editors, ISBN: 0-88318-802-3, 1988, 417 pp, Medical Physics Publishing, Madison (WI).

Topical

AAPM Report No. 10, A Standard Format for Digital Image Exchange, Task Force on Digital Image Data Exchange of the Science Council; ISBN: 0-88318-408-7, 1982, 11 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 30, E-Mail and Academic Computer Networks, Computer Committee Task Group 1, ISBN: 0-88318-806-6, 1990, 54 pp, Medical Physics Publishing, Madison (WI).

PACS: Basic Principles and Applications, Huang HK, ISBN: 0471253936, 1998, 519 pp, Wiley, New York.

PUBLIC EDUCATION

AAPM Report No. 53, Radiation Information for Hospital Personnel, Radiation Safety Committee; ISBN: 1-56396-480-5, 1995, 24 pp, Medical Physics Publishing, Madison (WI).

Cancer Patient's Guide to Radiation Therapy, Steeves R, ISBN: 0-944838-26-X, 1991, 87 pp, Medical Physics Publishing, Madison (WI).

Radiation and Health, Thormod Henrikson, H. David Maillie, David H. Maillie, ISBN: 0415271622, 2002, 240 pp, Taylor & Francis.

MEDICAL PHYSICS EDUCATIONAL AND PROFESSIONAL ISSUES

AAPM Monograph No. 27, Accreditation Programs and the Medical Physicist: 2001 AAPM Summer School Proceedings, Dixon R, Butler P, Sobol W, editors, ISBN: 1-930524-04-8, 2001, 364 pp. Medical Physics Publishing, Madison (WI).

AAPM Report No. 33, Staffing Levels and Responsibilities of Physicists in Diagnostic Radiology, Diagnostic X-Ray Imaging Committee Task Group 5; ISBN: 0-88318-913-5, 1991, 30 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 36, Essentials and Guidelines for Hospital Based Medical Physics Residency Training Programs, Presidential Ad Hoc Committee on Clinical Training of Radiological Physicists; ISBN: 1-56396-032-X, 1990, 147 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 38, The Role of a Physicist in Radiation Oncology, Professional Information and Clinical Relations Committee Task Group 1; ISBN: 1-56396-229-2, 1993, 12 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 42, The Role of the Clinical Medical Physicist in Diagnostic Radiology, Professional Information and Clinical Relations Committee; ISBN: 1-56396-311-6, 1994, 20 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 64, A Guide to the Teaching of Clinical Radiological Physics to Residents in Diagnostic and Therapeutic Radiology, Committee on the Training of Radiologists; ISBN: 0-944838-09-X, 1999, 32 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 79, Academic Program Recommendations for Graduate Degrees in Medical Physics, Education and Training of Medical Physicists Committee, ISBN: 1-888340-39-8, 2002, 72 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 80, The Solo Practice of Medical Physics in Radiation Oncology, Professional Information and Clinical Relations Committee Task Group 11, ISBN: 1-888340-41-X, 2003, 18 pp, Medical Physics Publishing, Madison (WI).

Current Regulatory Issues in Medical Physics, Martin M, Smathers J, ISBN: 0-944838-29-4, 1992, 458 pp, Medical Physics Publishing, Madison (WI).

Medical Physicists and Malpractice, Shalek R, Gooden D, ISBN: 0-944838-64-2, 1996, 140 pp, Medical Physics Publishing, Madison (WI).

RADIATION PHYSICS

AAPM Proceedings No. 8, Biophysical Aspects of Auger Processes, Howell RW, Narra V, Sastri K, Rap D, editors, ISBN: 1-56396-095-8, 1992, 418 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 37, Auger Electron Dosimetry (Reprinted from Medical Physics, Vol. 19, Issue 6), Nuclear Medicine Committee Task Group 6; ISBN: 1-56396-186-5, 1992, 25 pp, Medical Physics Publishing, Madison (WI).

AAPM Report No. 49, Dosimetry of Auger-Electron-Emitting Radionuclides (Reprinted from Medical Physics, Vol. 21, Issue 12), Nuclear Medicine Committee Task Group 6; ISBN: 1-56396-451-1, 1994, 15 pp, Medical Physics Publishing, Madison (WI).

Bluebells and Nuclear Energy, Reynolds A, ISBN: 0-944838-63-4, 1996, 302 pp, Medical Physics Publishing, Madison (WI).

Chernobyl Record: The Definitive History of the Chernobyl Catastrophe, Mould RF, 075030670x, 2000, 350 pp, IOPP, Bristol (UK).

My Life with Radiation, Ralph Lapp, ISBN: 0-944838-52-9, 1995, 168 pp, Medical Physics Publishing, Madison (WI).

Radiological Physicists, del Regato JA, ISBN: 0-88318-469-9, 1985, 188 pp, Medical Physics Publishing, Madison (WI).

Understanding Radiation, Wahlstrom B, ISBN: 0-944838-62-6, 1996, 120 pp, Medical Physics Publishing, Madison (WI).

Medical Physics Online Resources

Electronic Medical Physics World, International Organization for Medical Physics (http://www.medphysics.wisc.edu/~empw).

Medical Physics World, International Organization for Medical Physics (http://www.iomp.org); also available in hardcopy.

Global Online Medical Physics Book (http://www.iomp.org).

Organizations That Publish in Medical Physics

American Association of Physicists in Medicine (AAPM), One Physics Ellipse, College Park, MD 20740 (http://www.aapm.org).

American Institute of Physics, 2 Huntington Quadrangle, Melville, NY 11747-4502 (http://aip.org).

American Institute of Ultrasound in Medicine, 14750 Sweitzer Lane, Suite 100, Laurel, MD 20707 (http://www.aium.org).

American Roentgen Ray Society, 44211 Slatestone Court, Leesburg, VA 20176-5109 (http://www.arrs.org).

American Society for Therapeutic Radiology and Oncology, 12500 Fair Lakes Circle, Suite 375, Fairfax, VA 22033-3882 (http://www.astro.org).

British Institute of Radiology, 36 Portland Place, London, W1B 1AT, (http://www.bir.org.uk).

Health Physics Society, 1313 Dolley Madison Boulevard, Suite 402, McLean, Virginia 22101 (http://hps.org).

Institute of Physics and Engineering in Medicine (IPEM), Fairmount House, 230 Tadcaster Road, York, YO24 1ES, UK (http://www.ipem.ac.uk).

International Atomic Energy Agency, P.O. Box 100, Wagramer Strasse 5, A-1400 Vienna, Austria (http://www.iaea.org).

International Commission on Radiation Units and Measurements, Inc., ICRU, 7910 Woodmont Avenue, Suite 400, Bethesda, MD 20814-3095, USA (http://www.icru.org).

International Commission on Radiological Protection, SE-171 16 Stockholm Sweden (Elsevier).

International Society for Magnetic Resonance in Medicine, 2118 Milvia Street, Suite 201, Berkeley, CA 94704, USA (http://www.ismrm.org).

International Society for Optical Engineering (SPIE), PO Box 10, Bellingham WA 98227-0010 USA (http://spie.org).

National Council on Radiation Protection and Measurements, 7910 Woodmont Avenue, Suite 400, Bethesda, MD 20814-3095 (http://www.ncrponline. org).

Radiation Research Society, 10105 Cottesmore Court, Great Falls, VA 22066 (http://www.radres.org).

Radiological Society of North America, Inc., 820 Jorie Boulevard, Oak Brook, IL 60523-2251 (http://rsna.org).

Society of Nuclear Medicine, 1850 Samuel Morse Dr. Reston, VA 20190 (http://snm.org).

See also BIOMEDICAL ENGINEERING EDUCATION; MEDICAL EDUCATION, COMPUTERS IN; MEDICAL ENGINEERING SOCIETIES AND ORGANIZATIONS.

MEDICAL RECORDS, COMPUTERS IN

YESHWANTH SRINIVASAN
BRIAN NUTTER
Texas Tech University
Lubbock, Texas

INTRODUCTION

Computers play a vital role in almost every aspect of modern medical science. If we begin with a primitive definition of a computer, as a system with a processor unit capable of performing arithmetic and logical operations and a memory unit to store programs and data, then almost every sophisticated piece of modern medical equipment comes with a computer. Most standard medical procedures [e.g., magnetic resonance imaging (MRI), computed tomography (CT), positron emission tomography (PET), ultrasonography] use computers to record and to process data. Computer-based medical records would seem a natural choice for documentation of the recorded data. Moreover, the association of this data with computers presents numerous advantages including easy and random access to the data by multiple distributed users, globally searchable databases, and automated batch processing. However, the use of computer-based medical records is not as prevalent in today's medical community as one would expect. In this article, we explain the reasons for this anomaly, and we present methods to circumvent the problems associated with computer-based medical records and to allow medical practitioners to achieve the real potential of computer-based systems.

PAPER-BASED MEDICAL RECORD SYSTEMS

Medical records chronicle the transactions between healthcare professionals and their patients. They record every detail of a patient's visit to a hospital, clinic, or office from the time of the patient's arrival to the time the patient is discharged. Details recorded include the condition of the patient at the time of examination, procedures and medications administered to the patient, symptoms and observations, diagnoses and test results, together with the exact time of each relevant activity.

Traditionally, medical records have consisted of handwritten documents called *charts*. This remains the most prevalent form of recording patient information, especially in small hospitals and clinics. These charts typically record every piece of information that is judged potentially relevant to a patient's case. When a patient visits a hospital or a primary care physician for the first time, a new chart is made, and the patient name and contact information are recorded on it. When the patient is discharged from the hospital, a discharge note is made on the chart. During subsequent visits, new charts are added. All patient charts are maintained in a labeled file bearing the patient's name and other information to enable easy identification and retrieval.

However, paper-based records exhibit shortcomings in six important areas that are necessary to ensure adequate long-term care to patients and easy preservation of treatments administered.

1. *Accessibility*: Accessibility refers to the immediate availability of patient-related information at all times to all parties with a need to know. It is important to review a patient's medical history before commencing any new treatment. If a patient decides to visit a specialist, emergency room, clinic, or hospital other than the primary care physician, a request may be made for the patient's medical history to be faxed or couriered to the second facility. This information transfer may take between hours to days to complete, depending on the distance and arrangements between the two locations, available modes of communication, availability of authorized personnel to promptly respond to the request, and so on. During

emergencies, a delay of just a few minutes in administering the correct treatment to a patient could have very serious ramifications. A situation in which adverse reactions to particular medications known by a primary care physician simply are not made available in a timely fashion to emergency room personnel is all easy to imagine. The delay in disseminating accurate detailed background information is perhaps the biggest drawback of paper-based medical record systems.

2. *Legibility*: Paper-based records are generally handwritten documents, and one person's handwriting is often not readily understood by another. The fact that entries may be made by many different people makes understanding a chart even more difficult. Furthermore, paper-based records are subject to wear and tear, accidental immersion in water or other liquids due to leaks, misfortune or negligence, smudging of ink, and so on. These factors lead to reduced legibility of the entries on the charts and consequently make interpreting the charts more subjective.
3. *Bulkiness*: Paper-based records for a single patient initially consist of just a few charts, but they can quickly become a thick file with subsequent visits. Assuming that a small clinic sees ~ 10 new patients daily, that the physicians in the clinic take rotating holidays and that it is not a leap year, ~ 3650 new files will be added over a period of 1 year. The files of many of the existing patients also become bulkier and subsequently occupy more space. All of this paper forces the addition of storage space in the clinic for these new records, even though the space could often be better utilized for clinical purposes through the addition of much more profitable and useful treatment or diagnostic capabilities.
4. *Data Dimensionality*: Data dimensionality refers to the ability to interpret the available information in multiple, complementary fashions. The data contained in paper-based records simply provide information about the medical history of a particular patient. In order to obtain trends, distributions or statistical patterns about a particular patient or even groups of people in a geographical area, the required information has to be manually compiled from each record. This can be extremely laborious and time consuming. Furthermore, even a simple text search for a particular word or phrase can take impractically long if the search must be conducted visually through paper-based handwritten records.
5. *Integrity*: A patient will have multiple medical files, maintained simultaneously at every hospital or clinic ever visited, yet none of these records contain the complete medical history of the patient. Usually, the patient's primary care physician is the one who has the records that will provide the most accurate long-term information, and during times of emergency the primary care physician will frequently be the first one contacted for patient records. However, medical facilities rarely have in place any mechanism to update the primary care physician. If, for example, during a visit to a specialist it was found that a patient was allergic to a particular class of medications, the primary care physician's records will often not be updated to show this information. Even if the patient records from multiple sources were summoned, there will still remain a very realistic chance of missing one or more sources, leaving the completeness of paper-based records questionable. Paper-based systems also risk inadvertent loss or misfile of complete pages in copying, transporting, photocopying, faxing, and handling of records.
6. *Replication*: Paper-based medical records are prone to destruction by natural disasters, fires, leaks, and pests in storage areas. Because paper-based records are bulky, it is both expensive and tedious to replicate them and store back-up copies. Hence, if they are destroyed, patient medical histories can be completely lost.

Several research papers have been devoted to explaining the shortcomings of paper-based medical records (1,2). Nevertheless, paper-based records are still the most widely used method for maintaining patient information. The process of recording information in charts is a natural process of making notes, and it requires little specialized training. Once recorded, the information is relatively permanent and cannot be distorted easily. Furthermore, the chart often bears the signatures of both the patient and the appropriate medical personnel who treated the patient, which may be an important legal requirement. Such legal issues are a significant reason that many organizations resist wholesale changes to their paper-based record systems.

COMPUTER-BASED MEDICAL RECORD SYSTEMS

Paper-based records were the only method of storing patient medical information in use until ~ 25 years ago. Since that time, computer technology has experienced exponential growth, creating endless possibilities for developments within allied fields. The fields of medicine in general and medical records in particular have been no exception. Computer-based records were introduced to overcome the inherent limitations and problems of paper-based records and to alter patient medical record keeping in order to incorporate advancements in computer technology. The computer-based medical record is commonly given several names within differing environments, including Computer-based Patient Record (CPR), Electronic Health Record (EHR), and Electronic Medical Record (EMR). Throughout this article, we will use the name computer-based patient record and the abbreviation CPR to refer to computer-based medical records. The CPRs are fundamental to the role of computers in medical records, and a significant portion of this article deals with them.

A CPR is an electronic patient record that resides in a computer-based information system. Such systems are often specifically designed to augment medical practitioners by providing accessibility to complete and accurate long-term data, alerts, reminders, clinical decision support systems, links to medical knowledge databases, and other aids (3,4). The CPRs are essentially digital equivalents of paper-based records. Because this patient information is stored digitally, the enormous information processing and networking capabilities of computers can be utilized to drastically increase the efficacy of patient treatment. They easily overcome the shortcomings of paper-based records on the six key issues outlined in the previous section, and they provide several other fascinating features exclusive to them.

1. *Accessibility*: With the advent of the Internet and the World Wide Web, information on virtually any topic can be quickly and easily obtained. CPR databases can be stored in networked servers, which can be searched from any properly networked location. With properly designed databases and network systems, the complete medical history of a patient can be retrieved in a matter of seconds. This helps in ensuring the most prompt and accurate treatment possible, which is one of the fundamental reasons for maintaining a medical record.
2. *Legibility*: The information recorded on CPRs is presented using digitally reproduced fonts, which are then independent of the handwriting of the person entering the data into the system. Anytime patient information is required, it can be readily visualized on a viewing device, such as a CRT monitor or PDA, which does not introduce any wear and tear in the record itself. If a paper copy of a patient record is required for a particular activity, a new printout can be generated every time. Thus there is no subjective element in understanding the information.
3. *Bulkiness*: CPRs can be stored on a variety of media, including Compact Disc (CD), Digital Versatile Disc (DVD), and Hard Disk Drive (HDD). A single CD, 11.4 cm in diameter and weighing ~15 g, can store 700 megabytes (MB) of data. This translates into >100,000 pages of raw text or two hundred 500-page books, which would occupy >10 m of shelf space. The DVDs have >10 times the density of CDs, and HDDs offer >100 times the information content of a CD. This means that with proper database design, all of the medical records of all of 1 year's new patients at the aforementioned small clinic can now be stored on a single CD, and several years of patient data for all ongoing patient activities can be stored on just one DVD.
4. *Data Dimensionality*: Data on CPRs can be exploited and interpreted in many more ways than can a simple paper-based record. Entire databases can be subjected to automated search, and specific information from a single record or a collection of records can be retrieved in a matter of seconds. Trends and patterns in disease occurrence and treatment administration and effectiveness can readily be extracted and processed for limited or wide-ranging populations. Database Management System (DBMS) software can be used to plot the available information on graphs and pie charts, which greatly simplify the contents of medical records into a form that can be interpreted by laymen. All these operations can be done automatically using state-of-the-art software and require little, if any, human intervention. Techniques in data mining, for example, allow automated determination and analysis of data correlations among very large numbers of variables.
5. *Integrity*: CPRs are generally stored on local systems within hospitals, clinics, and physician offices, which means that medical records of a particular patient could be distributed across many systems throughout the world. Using modern networking techniques, these systems can be easily yet securely networked, and widely dispersed CPRs can be located and made available for inclusion in patient diagnosis and treatment. It would also be practical to develop a system in which all patient records are stored in a central repository, and any time new records of the patient are created or updated, they can be added to this central repository. Using either of these methods in a well-designed system, the complete medical history of a patient can be very rapidly retrieved from any of the authorized locations connected to the network.
6. *Replication*: Due to the high densities for computerized storage of data, CPRs occupy very little space when compared to paper-based systems. It is very easy and inexpensive to create back-up copies and to store these copies in geographically diverse locations. In a well-designed computer-based system with automated, networked backup, the complete medical history of every patient can still be retrieved even if a record storage facility is completely destroyed.

Exclusive Features of CPR

With considerable manual effort, the performance of paper-based records can be made comparable to that of CPRs within the six major requirements of a medical record outlined above. The efficiency advantages of computer-based medical record systems will nevertheless offer significant cost advantages over paper-based systems in achieving comparable performance levels. Significant improvements in patient treatment can result when CPRs are incorporated into expert systems and DBMS frameworks that engender an entirely new set of features for medical records.

1. *Content-Based Data Retrieval*: If a large hospital system is informed that a popular drug has been determined to cause serious side effects and that the manufacturer has consequently issued an immediate recall, the hospital system may hesitate to issue a broad public warning, because that action would lead to unnecessary panic and to patients flooding the system seeking confirmation on their prescriptions, even when their medications are completely unrelated to the recall. If the hospital maintained only paper-based records, it would take weeks to pore through patient charts to obtain the names

and locations of the patients to whom the drug was prescribed. Such a situation presents a perfect case for content-based data retrieval. With properly constructed CPRs, the same information can be much more rapidly retrieved through querying the database to fetch all patient records with the name of the recalled drug in the 'Drugs Prescribed' field. Then, only the patients who actually have been prescribed the recalled medication can be individually contacted with explicit instructions by a qualified staff member. This data search technique is known as Content-Based Data Retrieval (CBDR).

2. *Automatic Scheduling*: With CPR-based systems, routine periodic check-up or follow-up visits can be automatically scheduled, and e-mail reminders or automated telephone recordings can be sent directly to the patients. This automation will relieve healthcare professionals of a significant administrative burden.
3. *Knowledge-Based Systems*: Every physician attempts to solve a particular condition to the best of his limited knowledge (5). It is reported that there are $>5,000,000$ medical facts, which are constantly being appended, updated and questioned by over 30,000 medical journal publications each year (6). It is humanly impossible to remember all of these facts and to reproduce them accurately as needed. However, CPR databases all over the world can be rapidly and effectively searched in order to find information about patients who were previously diagnosed with the same condition, how they were treated and the end result. While the presiding doctor still makes the decisions and retains responsibility, he can have experimental evidence and experiential knowledge from other experts to assist his analysis (7). Diagnostic decision support systems like Quick Medical Reference (QMR) and Massachusetts General Hospital's DXplain provide instant access to knowledge bases of diseases, diagnoses, findings, disease associations and lab information.
4. *Expert Systems*: Human errors in recording information can be greatly reduced by programming the data-entry interface to reject anomalous values and to restrict the input to a finite set of possible choices. For example, a pulse rate of 300 beats min^{-1} would be highly unlikely. Potentially dangerous spelling mistakes in generating a prescription can be identified. Dates for follow-up activities can be checked against reasonable parameters. A well-designed interface would prompt a message to recheck such entries.
5. *In situ Data Access*: A medical records system in which a physician accesses a wireless PDA while conversing with a patient could allow direct bidirectional communication between a physician and a pharmacy, a laboratory, or a specialist. Collaborators on a complex case can simultaneously access each other's progress, even when the collaborators are geographically separated.

In 1991, the United States Institute Of Medicine (IOM) conducted a study on the advantages and disadvantages of CPRs and published its findings (3). The IOM concluded that CPRs greatly enhance the health of citizens and greatly reduce costs of care, and it called for widespread implementation of CPRs by 2001. However, 4 years past that deadline, widespread implementation still remains only a concept. Figures for CPR adoption rates in hospitals range from 3% to 21%, while CPR adoption rates for physician offices range from 20% to 25% (8). It is safe to say that these figures are not representative of the popularity that one would reasonably expect from the benefits of CPRs. The following section is devoted to explaining this anomaly.

ROADBLOCKS IN CPR IMPLEMENTATION

Recording and processing of information with CPRs requires installation of computer systems and software, and, in general, more training of healthcare professionals is required to get them acquainted with CPR systems than with paper-based record systems. Furthermore, realizing the full potential of CPR systems requires not only that patient records issued in the future should be CPRs but also that previously taken paper-based records should be converted into CPRs. Some of the important factors affecting this onerous large-scale conversion will now be discussed.

1. *Cost*: Expense is usually the major obstacle to the full-fledged incorporation of CPRs. Depending on whether the application is for the office of a single physician or for a large hospital, the basic infrastructure needed to establish a CPR system could vary from a single PC to a complete network of workstations and servers loaded with expensive network, database and records management software. Both hardware and software computer technologies become outmoded very quickly, and these technologies require constant update, increasing maintenance and operation costs. There is also a significant cost factor involved in hiring data-entry operators to convert existing paper-based records to CPRs and in training healthcare professionals to then operate the new systems and software. Although several studies have shown that CPRs are effective in the long run (9,10), both hospitals and physician offices remain apprehensive about investing many thousands of dollars on a new technology while the tried and tested paper-based records work reasonably effectively.
2. *Abundance of Choices and Lack of Standards*: There are >100 CPR management software systems available in the market today, and additional entries arrive in the marketplace annually. Each software approach will have its advantages and disadvantages, and it is never easy to decide which one is best suited for a particular application. Furthermore, different software interfaces use different templates for the patient record, recording slightly different patient data within very different structures. A

standardized CPR format independent of the specific hardware and software running on a particular system is yet to be developed. This lack of standardization in CPR formats severely cripples cross-system interaction and comprehensive integration of CPR systems. The marketplace has, to date, failed to produce a clear market leader or an industrial body with the ability to enforce such standards. A customer contemplating a purchase and the subsequent investments in training and maintenance may have serious reservations concerning the ability to change from the products of one manufacturer to another.

3. *Confidentiality*: The United States government passed the Health Insurance Portability and Accountability Act (HIPAA) in 1996, protecting a patient's right to confidentiality and specifying the information on a medical record that can be shared and the information that must be considered confidential. A record can be made available only on a "need to know" basis, wherein only personnel who absolutely need to know the data are granted access. For example, physicians participating in the diagnosis and treatment process of a specific patient are granted complete access to the entire medical record, while health agencies involved in conducting demographic studies are granted access only to information that does not and can not reveal the identity of the patient.
4. *Security*: In the case of paper-based records, it is relatively easy to control access and to ensure security by keeping those records under lock and key. The security of CPRs will be constantly threatened by hackers, trying to break into a system in order to retrieve confidential patient information. Unless properly protected, CPRs are also vulnerable to being compromised during transmission across a network. Unlike paper-based records, CPRs can be easily manipulated, intentionally or unintentionally, in fashions that do not appear obvious. These vulnerabilities represent an enormous security problem, and effective solutions are required before CPRs can be put to widespread use. Furthermore, efforts to provide network security have significant costs, requiring manpower, equipment, and frequent software upgrades.
5. *Apprehension Among Patients*: Even in developed countries like the United States, considerable uneasiness exists among patients about the use of computers to maintain patient medical records. While some of this phenomenon can be ascribed to the reluctance of the older generation in accepting new technology, the most important reason for this apprehension is the lack of clear understanding by patients about the benefits of CPRs. One of the easiest ways to overcome this fear is for healthcare professionals to directly reach out to the patients and explain the long-term advantages of CPRs.
6. *Legal Issues*: The validity of CPRs in a court of law is still questionable. The Millennium Digital Commerce Act, otherwise known as the Electronic Signature Act, was passed in 2000 (although the portion relating to records retention was passed in 2001). The act established that the electronic signature and electronic records cannot be denied legal effect simply because they are electronic and are not signed in ink on paper. However, legal implications in CPR transactions extend far beyond this act. For example, if it is found that the software system used to manage CPRs does not comply with stipulated security requirements, the validity of a CPR can be questioned in court, even if no indication of tampering can be identified. The CPR systems are also prone to computer viruses, worms and Trojans, which exploit loopholes in the computer security defenses to gain unauthorized access to sensitive data and/or to maliciously corrupt the information.

Despite significant progress toward overcoming these hurdles, complete transition from paper-based records to CPRs is a time-consuming and capital-intensive process, and during the midst of this transition, the benefits are unlikely to be immediately apparent. In the following section, some of the options available to healthcare professionals in implementing a CPR system, and several technologies available to ensure secure storage and transmission of confidential medical information are reviewed.

FEATURES OF CPR SYSTEMS

A complete CPR system can consist of individual workstations, data input devices (keyboards, mice, light pens, scanners, cameras), output devices (printers, plotters, and display monitors), Database Management System (DBMS) software, text, image and video processing hardware and software, and the networking infrastructure for intra- and intersystem communication. The security and reliability requirements of these systems should never be understated. With so many factors involved, there is always room for improvement, and a universally perfect combination has yet to be implemented. Nevertheless, a wide variety of CPR systems have been successfully implemented, tested and put to everyday use, with most such systems working successfully in the environment for which they were designed. We will now explore some of the general features of CPR systems and their components, and we will present advantages and disadvantages of the choices available.

System Architecture

A fundamental decision that must be made before implementing a CPR system is the choice of the system architecture to be used. There are two broad choices, which we now discuss.

1. *Centralized Architecture*: A block schematic of a centralized architecture as implemented in CPR systems is shown in Fig. 1. A CPR system based on the centralized architecture has a central server that contains complete medical records of all patients in the system. Any modification to a record must be

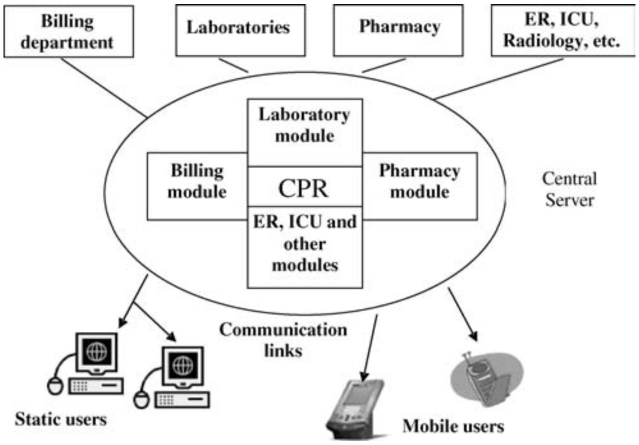


Figure 1. Centralized Architecture (Reprinted with permission from (The Computer-Based Patient Record: An Essential Technology for Health Care, Revised Edition) © (1997) by the National Academy of Sciences, courtesy of the National Academies Press, Washington, D.C.)

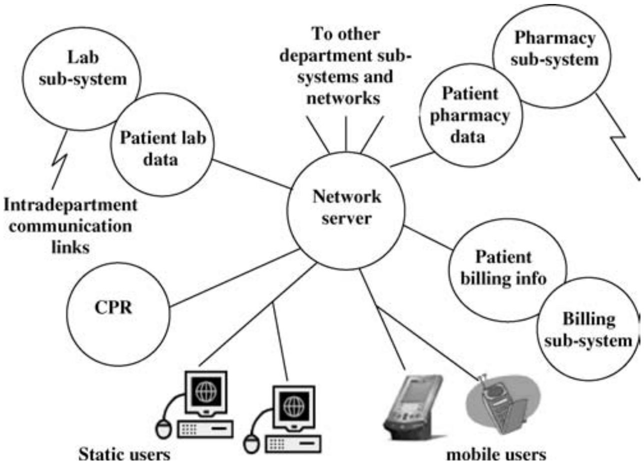


Figure 2. Distributed Architecture (Reprinted with permission from (The Computer-Based Patient Record: An Essential Technology for Health Care, Revised Edition) © (1997) by the National Academy of Sciences, courtesy of the National Academies Press, Washington, D.C.)

done to the CPR stored in this central repository to have any lasting effect. Communication links may be provided to other departments that will immediately communicate any new information to the server, thereby keeping the CPR on the central server fully updated. The principal advantage of this architecture lies in the immediate availability of the complete record for use elsewhere within the system. It is also easier to manage and to maintain the resources of the system, because they are located in one single location. However, a poorly designed system can be prone to server overloading, leading to delayed responses. Server or network failures can bring the system to a complete standstill, and if the data on the server is corrupted or destroyed, the medical record activity subsequent to system backup may be completely lost. These issues can be resolved, but the solutions can be expensive.

2. *Distributed Architecture*: The alternative to the centralized approach is the distributed architecture. A block schematic is shown in Fig. 2. In this approach, the load of the server is distributed among smaller subsystems, distributed through many departments. No department contains the complete medical record of all patients. The records are completed “on demand”, that is, when a user at a system workstation must obtain complete information, the workstation sends requests to each individual subsystem, receives data back from them, and arranges them into a record. This system continues to function even if one or more of the subsystems become inoperative. However, the response time, defined as the time elapsed between the user request to fetch a record and the arrival of the complete record back to the requesting workstation, might become significant if even one of the subsystems is overloaded. The architecture also provides more opportunities for security breaches than the centralized approach. There may also be difficulties related to each subsystem using different data structures to manage and to store records, making the design of an effective DBMS potentially quite complex. These issues have been successfully addressed in specific implementations of distributed CPR systems.

DBMS and Media Processing

The general components of a CPR are shown in Fig. 3. An ideal DBMS would be able to seamlessly integrate different media formats (text, images, three-dimensional (3D) models, audio and video). Four important types of DBMS commercially available are hierarchical, relational, text- and object oriented. Each of these is better suited for one or more different media and system architectures than the others. Many modern CPR systems, especially those based on a distributed architecture, use a combination of commercial DBMSs or utilize a custom DBMS.

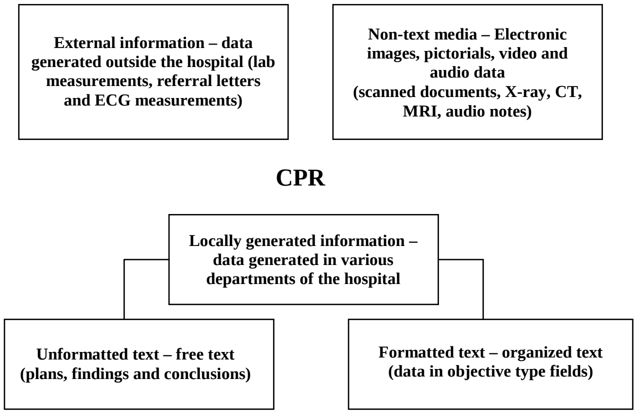

External information – data generated outside the hospital (lab measurements, referral letters and ECG measurements)
Non-text media – Electronic images, pictorials, video and audio data (scanned documents, X-ray, CT, MRI, audio notes)
CPR
Locally generated information – data generated in various departments of the hospital
Unformatted text – free text (plans, findings and conclusions)
Formatted text – organized text (data in objective type fields)

Figure 3. Components of a CPR.

An important aspect of a DBMS is the mechanism that it uses for data retrieval. The details of record construction can play a significant role in creation of databases that can be efficiently and accurately searched. Entries in formatted fields can be searched much more easily than scanned documents, images or nonformatted text. A DBMS should also provide organization of the data in a manner that is sensible and easy to comprehend. Well-organized databases will prove easier to manage and faster to search.

Security and Legal Issues

A core issue in the design of CPR systems is security. A single security breach could be used to access thousands of complete CPRs, which are then vulnerable to many malicious activities. Thus, it is extremely critical to protect individual CPRs from unauthorized access, as well as to protect the CPR system itself from hackers attempting to access the system or attempting to cause it to fail. In addition, the CPR data must be protected before it can be sent across a network, due to potential communication leaks that can land sensitive information in the wrong hands. Tools such as network packet sniffers are extremely powerful if an unauthorized party can achieve physical access to network interconnection hardware or cabling. In general, five conditions must be satisfied before patients or health care personnel are granted access to the records on a network (11):

1. The patient must not have recorded an explicit prohibition of CPR transmission across the network.
2. The patient must have consented to the transfer of the partial or complete CPR to the recipient institution.
3. The identity of the patient must be accurately verified and authenticated.
4. The identities of the healthcare provider and the recipient institution must be accurately verified.
5. It must be verified that the patient is mentally competent. If the patient is found to be mentally incompetent, then the need for the patient record must be sufficiently urgent that if it is not made available immediately, serious harm could result to the patient's health.

For a CPR produced by a CPR system to be a legally valid document, the CPR should be able to answer the following questions (12).

1. *Who Wrote It*? A single CPR is frequently handled by many different people, with several healthcare providers entering, updating, or removing information. A CPR must include the identities of all the people who have modified the record, together with the information that they have modified. This involves secure access control using smart cards, PINs, fingerprints, passwords, etc. to verify the identity of requestors before they will be allowed to access or to modify the information on the records.
2. *When Was It Written*? Time and date information is maintained by automatically attaching timestamps to the CPRs. The system must be designed to ensure that unauthorized personnel cannot change the system time or any timestamp in any CPR.
3. *What Does It Say*? Computer technology is reinvented roughly every 5 years. Data storage and retrieval media become obsolete very quickly. For example, the magnetic tapes that were very popular 25 years ago can be found only in museums today. It is thus important for CPR systems to incorporate these technological developments, and yet not become outmoded too quickly, so that the contents of current CPRs and archived CPRs can both be read and processed to obtain the same information.
4. *Has the Information Been Altered*? Undocumented alteration of information will prove to be the most important legal issue, and hence, the most difficult one to solve. A CPR system must be designed to allow determination of whether the information is indeed as originally recorded or whether it has been subsequently modified in a manner that would affect possible legal proceedings. Several effective solutions, based on cryptographic techniques, exist.

A very reliable solution to ensure that a particular CPR provides answers to all of these four questions is to attach digital signatures to them. Just as an ink-based signature provides legal acceptability for paper-based records, digital signatures can ensure the legal acceptability of a CPR. Digital signatures offer both signatory and document authentication, and they have been proven to be more effective than the ink-based signature (13). Signer authentication provides the capability to identify the person who signed the document, while document authentication offers the capability to determine whether a document was altered after it was signed.

Digital signatures use public-key encryption schemes to provide this functionality. In public-key encryption schemes (14), every user gets a pair of *keys*: a public key, which everybody is allowed to know, and a private key, which is kept as a secret known only by the individual and the authority that issues and verifies the keys. A digital signature is computed using the data bytes of the document and both the private and public keys and attached to the document. Consider that person A creates a CPR and attaches a digital signature to it using his public and private keys. Anybody with access to A's public key can verify that the CPR was actually created by A using the CPR, the digital signature and A's public key. If the result is correct, according to a straightforward mathematical relationship, A is authenticated as the signatory of the CPR, and the CPR is authenticated as accurate. Any alteration of the CPR after the calculation and attachment of the digital signature would corrupt the digital signature and would then cause the CPR to be identifiable as modified. In this case, the CPR would be considered fraudulent, although the source of the fraud would not necessarily be identifiable.

Public-key encryption can also be used to secure CPRs before transmitting them across a network. Consider that person A desires to send a CPR to person B. Person A looks up the public key of B and encrypts the CPR using B's public key. The CPR can only be decrypted using B's private key. Since only B has access to his private key, only he can decrypt the CPR. The principle of public-key encryption can be extended to CPRs to protect the authenticity of the information, to identify malicious alterations of the CPR, and to secure transmission across a network.

POINTERS TO THE FUTURE

Much of this article has been devoted to explaining the benefits of the CPRs and the factors hindering their widespread implementation. Despite their tremendous potential, the development of commercial CPR systems has not reflected the progress made in many related fields in computer technology. In this section, some of these related technologies that, in the future, the authors feel will have great potential to broaden the range of advantages of computer-based medical record services and to make them sufficiently secure are presented.

Radio Frequency Identification Tags

Radio Frequency (RF) identification (RFID) refers to an automatic ID system that uses small RF devices to identify and to track individual people, pets, livestock, and commercial products. These systems are used to automatically collect highway tolls and to control access to buildings, offices and other nonpublic sites. An RFID tagging system includes the tags themselves, a device to write information to the tags, one or more devices to read the data from the tags, and a computer system with appropriate software to collect and to process information from the tag. Many applications can use the same devices to both read and write the RFID tags. While the physical layout of the tag may vary according to the application, its basic components will include an intelligent controller chip with some memory and an antenna to transmit and receive information. According to its power requirements, an RFID tag will be classified into one of two types. Active tags have their own power source and hence provide greater range. Passive tags will be powered by RF pulses from the tag reader (or writer) and thus exhibit no shelf life issues due to battery exhaustion.

While plans are underway to replace bar codes with RFID tags on virtually every single commercial product, the field of medical informatics has found some unique applications for RFID tags. The U.S. Food and Drug Administration (FDA) recently approved the implantation of RFID tags on humans (15). These tags are similar to the tags being used on animals, and they are implanted under the skin in the triceps area. The chip contains a 16-digit number that can be readily traced back to a database containing the patient's information. One such chip made by VeriChip costs $\sim$ \$125, exclusive of the cost of implantation. This technology is expected to be a boon to individuals with life-threatening medical conditions and to lower medical costs by reducing errors in medical treatment.

Testing of replacement of bar codes with RFID tags in patient bracelets is ongoing. Unlike bar codes, RFID tags do not require clear line of sight between the bar code and the bar code reader, nor do they require active operator intervention. This ensures that healthcare workers will not fail to scan a patient ID bracelet. One such successfully tested system is Exavera's eShepherd, which combines RFID tags with Wireless Fidelity (Wi-Fi) networks and Voice over IP (VoIP) to implement a single system that will track patients, staff and hospital assets (16). The Wi-Fi routers can deliver patient information directly to a physician's handheld PDA every time any RFID transceiver detects a patient. Physicians and hospital staff can have the patient information whenever and wherever they want, and they do not have to refer repeatedly to a secure physical filing area to retrieve patient records.

A major factor impeding widespread implementation of RFID tags is the legal and ethical issue of keeping such detailed records of people, their activities, and all the things they buy and use, without their consent. However, once this issue has been resolved, RFID tags will change the way patient records are processed.

Biometrics

Biometrics refers to automatic recognition of people based on their distinctive anatomical and behavioral characteristics including facial structure, fingerprint, iris, retina, DNA, hand geometry, signature, gait and even the chemical composition of sweat (17). Biometric systems have been used in wide ranging applications like user authentication before entry into buildings and offices, criminal investigation, and identification of human remains.

Biometric systems will also find an excellent application in management of medical records. Individual traits of people who are authorized to access the record can be stored along with the record. When the system receives a request from a user to retrieve the record, the system will attempt to match the biometrics of the user to the database of biometrics of all the people who are authorized to access the record. The record is fetched only if the match is positive. In this manner, a CPR can be viewed and modified only by authorized personnel. This is potentially an effective application for biometric systems, because the system can be trained with as many samples as needed from authorized personnel. Biometric system applications where a strong need exists to verify the authenticity of claimed membership in a preidentified population have been successfully tested with almost 100% accuracy. Such systems also offer a deterrent to would-be intruders because the biometric information associated with an unsuccessful access attempt can be retained and used to identify the intruder.

Generally, the biometric information of a user is encoded on a smart card that the user must use to gain access into the system. This provides a far better authentication solution than using just a user name and a password, because both of these can be forged fairly easily. Most current biometric systems are based on the uniqueness of human fingerprints, and systems based on more complicated biometric features are currently undergoing investigation. The development of superior biometric systems holds the key to more secure CPR authentication.

Content-Based Image Retrieval

Content-Based Image Retrieval (CBIR) is the process of retrieving images from a database based on the degree of similarity in content between an example image (the query image) and the various images contained in the database. Traditional non-CBIR data retrieval mechanisms apply text-based methods, in which a string query is matched to user-generated descriptions of documents or to contents of specific fields in a record. When patient images are integrated into patient records, simple text-based searches will not serve the same purpose. Text descriptions of images are very subjective and require data entry by a knowledgeable user. If that same image is to be added to a different database, the accompanying description also must be added if the image is to be retrievable.

These techniques utilize both global properties, such as image statistics and color distributions, and local properties, such as position and texture, to retrieve images that appear "similar" to the query image. This is a powerful tool for physicians, because CBIR enables them to compare the image of a current patient (query image) to similar images of other patients in the database, to examine descriptions attached to those images in their corresponding CPRs, and to study the treatment administered in those cases, all in a matter of seconds.

More and more frequently, scanned documents such as referral letters and scanned paper-based records are being added to CPRs. These documents can be treated as images, and CBIR systems based on character recognition technologies can search for text descriptions on these documents, making these documents compatible with text-based searches. This technology can save a great deal of time, money and errors in converting historical or ongoing paper-based records to CPRs, because the records can simply be scanned and need not be entered again into the CPR system. As more and more visual aids are added to medical diagnostics, CBIR will become indispensable to CPR management.

Steganography and Watermarking

Steganography and cryptography are related but distinct methods that are used to ensure secure transmission and secure archival of information. In cryptography, messages are encrypted in such a manner that unauthorized recipients may not decrypt the messages easily, using long, secure passwords and complex mathematical algorithms to drastically alter the data. Often, however, the encrypted messages themselves may be obtained fairly easily, because they are transmitted over insecure networks and archived on insecure servers. In steganography, the very presence of secure information is masked by hiding that information inside a much larger block of data, typically an image.

If an oncologist wishes to get a second opinion from a gynecological specialist concerning a diagnosis of cervical cancer, he might elect to send a digital image of the cervix of the patient. He would also include his diagnosis, relevant patient details including other complications the patient may have, and a referral letter, all of which are sensitive and confidential information. The traditional method to send the supplemental information electronically would be to encrypt the information using public-key encryption techniques and to transmit it as separate files together with the image. If the transmission is intercepted by a malicious party seeking private data, the transmission would garner interest, because it would be quite obvious that any encrypted information following the image information would likely be related to that image. The data thief must still work to decrypt the transmission and to reveal the confidential information. Decryption itself is difficult without extensive computational capacity, but data thieves will often find getting the necessary password to be a much easier method to access the data. The disadvantage of encryption is that the data thief knows what to try.

An alternate data transmission method would use steganography. The sensitive information can be embedded in inconspicuous areas of the image (18). Thus, no additional information or files are transmitted, and the image will raise less suspicion that associated sensitive data can be located. Although the embedded data is computationally easier to decode than the encrypted messages, it will only be decoded by people with foreknowledge that there is information embedded in the image. Furthermore, the embedded data can be encrypted before it is embedded, so that the data thief has another level of barrier. The information is much more difficult to locate, and the thief still has to decrypt or to steal the password.

Usually, the clinically important sections of the image will be identified, and only the remaining regions will be used for embedding the secret information. Several sophisticated steganography techniques, like BPCS, ABCDE, and modified ABCDE (19–21) have been successfully implemented. These methods can hide significant amounts of information without clinically altering the image, although hiding too much data leaves evidence that can be detected mathematically using tools from the field of steganalysis. Figure 4 shows an example of data hiding using the

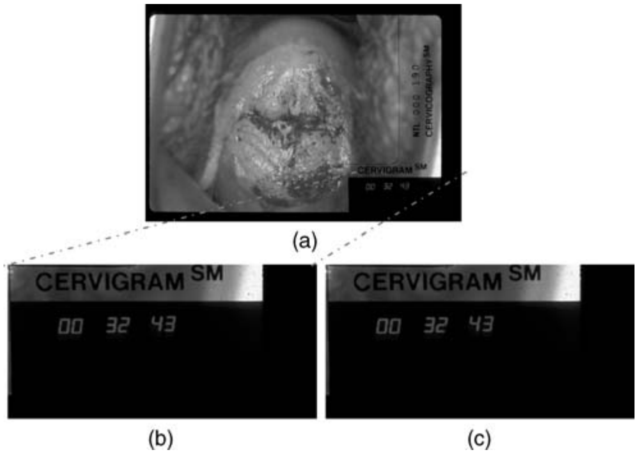


Figure 4. Steganography. (a) Original digital image of the cervix. (b) Section of clinically unimportant segment of image in a. (c) Section in b encoded with secret information.

modified ABCDE scheme. Figure 4a is a digital image of a cervix, and Fig. 4b shows the bottom right section of this image, which contains no clinically important information. This section forms an 864×432 RGB image, which requires 1,119,744 bytes to represent directly or 295,727 bytes to represent when losslessly compressed to PNG format. Figure 4c is the section in Fig. 4b after 216,326 bytes of sensitive information have been embedded. This hidden data corresponds to >40 pages of pure text, at 5000 characters per page. Figure 4c visually appears very similar to the section in Fig. 4b, even with this extensive additional data embedded. An even more secure solution would encrypt the sensitive information before embedding it inside the carrier image. This way, even if steganalysis methods detected that the carrier image contains embedded information, the decryption problem still remains.

Watermarking techniques are similar to digital signatures in the sense that they can provide owner and document authentication for digital media. They are usually nonvisibly detectable signals embedded into the media that can be checked, by authorized personnel, to verify the validity of the media and to trace its copyright holders. Watermarking and steganography techniques can also be used to identify times, dates, locations, and information that could be used to cross-reference to a patient, so that misidentification or misfiling of image data under the wrong patient file can be detected and corrected. It is the opinion of the authors that both these technologies will provide valuable solutions for secure CPR systems.

CONCLUDING REMARKS

This article attempted to provide a brief overview of the applications of computers in medical records, the benefits of computerization, and issues concerning widespread deployment of computer-based medical record systems. The computerization of medical records is a vast area of research, employing many people from diverse fields of medicine, science and engineering. While we have attempted to summarize and present the material from our own perspective, the topic itself is explored in great detail by several worthy publications. There are literally thousands of publications that attempt to explain or to solve one or more of the issues presented in this article and to provide much greater detail than is possible in this article. Thus, the authors felt it appropriate to leave the reader with a few resources, both in print and on the World Wide Web (WWW), to obtain more information about computers in medical records.

Much of the material in this article has been inspired from the extensive discussions in Refs. 3 and 22. These books provide interesting observations and refer to a wealth of references that pioneered the computer revolution in the medical field. Two more recent books, (23 and 24), also provide excellent insight into the various aspects of using computers to manage medical records. The WWW contains many websites that provide solid insight into various aspects of medical records management. Ref. 6 is one such site that is constantly updated with information about upgrading to CPR systems, cost benefit calculators, software vendor surveys, and from basic tutorials on every aspect of CPR management. Refs. 25 and 26 provide excellent evaluations of commercially available software for CPR management.

ACKNOWLEDGMENT

The authors would like to thank Mr. Rodney Long at the U.S. National Library of Medicine, Bethesda, Maryland, for providing the images of the cervix used in the example on steganography.

BIBLIOGRAPHY

1. Shortliffe EH. The evolution of electronic medical records. Acad Med 1999;74(4):414–419.
2. Burnum JF. The misinformation era: The fall of the medical record. Ann Intern Med 1989;110:482–484.
3. Dick RS, Steen EB, Detmer DE, editors. Institute of Medicine. The Computer-Based Patient Record—An Essential Technology for Health Care. Washington (DC): National Academy Press; 1997.
4. Shortliffe EH, Perreault LE, Fagan LM, Wiederhold G. Medical Informatics Computer Applications in Health Care and Biomedicine. 2nd ed. New York: Springer-Verlag; 2000.
5. Covell DG, Uman GC, Manning PR. Information needs in office practice: Are they being met? Ann Intern Med 1985; 103:596–599.
6. Voelker KG. (None). Electronic Medical Records [online]. http://www.emrupdate.com. Accessed 2005, April 10.
7. Weed LL. Medical records that guide and teach. New Engl J Med 278(11):593–600 and 278(12):652–657.
8. Brailer DJ, Terasawa EL. Use and adoption of computer-based patient records, California Healthcare Foundation report, Oct. 2003, ISBN 1-932064-54-0.
9. Wang SJ, et al. A cost-benefit analysis of electronic medical records in primary care. Am J Med Apr 1 2003;114(5):397–403.
10. Classen DC, Pestonik SL, Evans RS, Burke JP. Computerized surveillance of adverse drug events in hospital patients. J Am Med Assoc 1991;266:2847–2851.
11. David MR, et al. Maintaining the confidentiality of medical records shared over the Internet and the World Wide Web. Ann Intern Med 1992;127(2):138–141.
12. Cheong I. The legal acceptability of an electronic medical record. Aust Fam Phys Jan. 1997;26(1).
13. Askew RA. Understanding Electronic Signatures [online]. Real Legal. Available at http://www.reallegal.com/downloads/pdf/ESigAskewWhitePaper.pdf. Accessed 2005. April 10.
14. Dent AW, Mitchell CJ. User's Guide to Cryptography and Standards. Boston: Artech House; 2004.
15. Sullivan L. FDA approves RFID tags for humans. Inform Week Oct. 2004.
16. Collins J. RFID remedy for medical errors. RFID J May 2004.
17. Jain AK, et al. Biometric: A grand challenge. IEEE Conference on Pattern Recognition, Vol. 2; Aug. 2004. pp 935–26.
18. Johnson NF, Duric Z, Jajodia S. Information hiding: Steganography and Watermarking—Attacks and Countermeasures. Norwell (MA): Kluwer Academic; 2001.
19. Kawaguchi E, Eason RO. Principle and Applications of BPCS-Steganography. Proc SPIE Int Symp Voice, Video Data Communications; 1998.
20. Hirohisa H. A Data Embedding Method Using BPCS Principle With New Complexity Measures. Proc Pacific Rim Workshop on Digital Steganography. July 2002; p 30–47.

21. Srinivasan Y, et al. Secure Transmission of Medical Records using High Capacity Steganography. Proc IEEE Conf Comput.-Based Medical Systems; 2004. p 122–127.
22. Dick RS, Steen EB, editors. Institute of Medicine, The Computer-Based Patient Record—An Essential Technology for Health Care. Washington (DC): National Academy Press; 1991.
23. Hartley CP, Jones III ED. EHR Implementation: A Step-by Step Guide for the Medical Practice. Am Med Assoc Feb. 2005.
24. Carter JH. Electronic Medical Records: A Guide for Clinicians and Administrators. American College of Physicians March 2001.
25. Rehm S, Kraft S. Electronic medical records—The FPM vendor survey, Family Practice Management. Am Acad Family Phys Jan. 2001.
26. Anderson MR. EMR frontrunners, Healthcare informatics online, May 2003.

See also EQUIPMENT ACQUISITION; OFFICE AUTOMATION SYSTEMS; PICTURE ARCHIVING AND COMMUNICATION SYSTEMS; RADIOLOGY INFORMATION SYSTEMS.

MICROARRAYS

NEIL WINEGARDEN
University Health Network
Microarray Centre, Toronto
Ontario, Canada

INTRODUCTION

Microarrays allow for the simultaneous, parallel, interrogation of multiple biological analytes. Originally, microarrays were devised as a method by which gene expression could be measured in a massively parallel manner (all the genes in the genome at once), however, recent advances have demonstrated that microarrays can be used to interrogate epigenetic phenomena, promoter binding, protein expression, and protein binding among other processes. The overall process is reliant upon the manufacture of a highly ordered array of biological molecules, which are typically known entities. The features of this array behave as probes, which react with and bind to the unknown, but complimentary material present in a biological sample. Here we will focus specifically on gene expression (deoxyribonucleic acid, DNA) microarrays, which can be used to assay the activity of thousands of genes at a time.

In 1993, Affymetrix published a novel method of using light directed synthesis to build oligonucletide arrays that could be used for a variety of biological applications (1). Shortly thereafter, a group lead by Patrick Brown and Ron Davis at Stanford University demonstrated that robotically printed cDNA arrays could be used to assay gene expression (2). Now, more than a decade after this initial work was made public, both types of DNA array are commonly found in genomics laboratories.

BASIC PRINCIPLES

A DNA microarray contains a highly ordered arrangement (array) of several discrete probe molecules. Generally, the identity of these probes, be they cDNA or oligonucleotides, is either known or can be determined readily. The probes are deposited by some means (see the section Fabrication of Microarrays) onto a solid-support substrate such as glass or silicon. DNA microarrays take advantage of a basic characteristic of DNA, namely, the ability of one strand of DNA to find its complementary strand in solution and bind (hybridize) to it. This hybridization event is highly specific following standard Watson–Crick base pairing rules (Fig. 1).

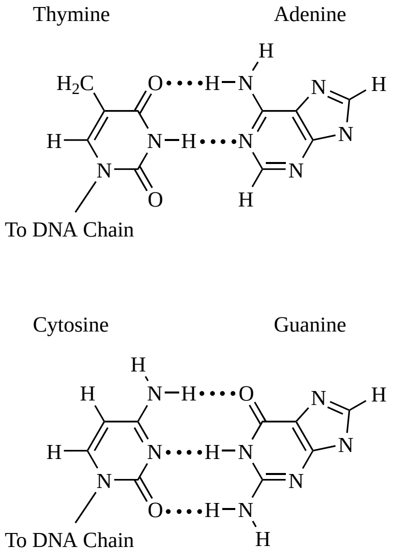


Figure 1. Watson–Crick base pairing interactions. During hybridization, specific base-paring interactions occur by which Thymine (T) binds specificly to Adenine (A) and Cytosine (C) binds specifically to Guanine (G). The binding of these bases to one another is mediated by hydrogen bonding as shown. The GC base pairs are stronger by virtue of the three hydrogen bonds formed compard to only two for AT.

Gene Expression

With some exceptions, the genetic makeup of every cell in an organism is the same. Each cell has the same complement of genes, which comprise the organism's genome. The subset of genes that are active in a particular cell dictate that cell's function. When we say a gene is active or *expressed*, we mean that particular gene is being transcribed. Transcription is the process by which ribonucleic acid (RNA) polymerase II (an enzymatic complex) reads a gene and creates a complementary copy of messenger RNA (mRNA). The more a gene is transcribed, the more copies of mRNA will be present in a cell. Thus genes that are highly active in the cell will be represented by multiple copies of mRNA, whereas genes that are inactive in the cell will have very few or no copies of mRNA in the cell. Microarrays function to measure the amount of mRNA present in the cells of a biological sample such as a tumor biopsy. The activity of the genes is inferred from this measure.

Gene Structure

In higher eukaryotes, somatic cells (diploid) have two copies of every gene: one maternally and the other paternally derived. In the context of the diploid cell, each copy is termed an allele. In the case where both inherited alleles are the same for a given gene, that gene is said to be homozygous. If the two alleles are different, then the gene is heterozygous. Alleles may be either dominant (phenotypically manifested regardless of what the other allele is), or recessive (phenotypically manifested only in the absence of a dominant allele). In the case of a heterozygous gene, the dominant allele will be phenotypically manifested and the recessive allele will not. If both alleles are different, but dominant, they are termed codominant and both alleles will elicit a phenotype. The gene is comprised of DNA, which is double stranded. One strand is the sense strand or the strand that encodes the information, which will be ultimately represented in mRNA. The other strand is said to be anti-sense and is the strand of DNA that is actually read by the RNA polymerase to generate the mRNA. DNA has directionality: A gene is transcribed starting at the 3′ end of the antisense strand of the DNA and is read toward the 5′ end. The resultant mRNA is made from the 5′ to the 3′ end.

Genes are regulated by specific sequences of DNA that lie outside the coding region of the gene. The first such sequence is the promoter. Promoters bind the transcriptional machinery (RNA polymerase II) that performs transcription. Promoters are found 5′ (upstream) of the gene and are proximal to the transcription start site. An additional class of regulatory sequence called an enhancer may be associated with the gene. Enhancers may lie upstream, downstream, or internal (usually in noncoding regions termed introns) to the gene (3). Specific transcription factors bind enhancers and promote recruitment or activation of the basal transcriptional machinery. It is the coordinated function of the promoter and enhancer, with the transcription factors that bind them, that control if a gene is active or not within the cell. Thus, genes are regulated, and can be turned on, off, or modulated up or down by the regulatory mechanisms of the cell.

RNA Isolation

Ribonucleic acid must be isolated from cells in order to prepare the material for hybridization to the array. A cell contains three major species of RNA: mRNA, transfer RNA (tRNA), and ribosomal RNA (rRNA). Together they are refered to as total RNA. For the purpose of gene expression experiments with microarrays, the mRNA is the species we are interested in and represents $\sim 1\%$ of total RNA. In order to isolate total RNA from cells, one of two main modalities is used: solution- or solid-phase extraction. In solution-phase methods, cells are lysed in the presence of isothiocyanate in order to inactivate any RNases (naturally occurring enzymes that nonspecifically degrade RNA). The lysate is then extracted with an acidified phenol: chlorophorm:isoamyl alcohol solution. The RNA selectively partitions to the aqueous phase of this mixture away from proteins and DNA. The aqueous phase is removed and RNA is precipitated out of solution using isopropyl alcohol at high salt concentrations. Solid-phase methods make use of the variable binding activity of RNA to a silica matrix at high and low salt conditions. Cells are again lysed in the presence of isothiocyanate. The high concentration of isothiocyante used in this methodology not only inactivates the RNases, it also selectively precipitates proteins out of solution. The lysate is applied to a column containing a silica filter at the bottom. The lysate is pulled through the column via vacuum or centrifugation, thereby removing the proteins and cellular debris. In this method, DNA may also bind to the column, and as such contaminating DNA is removed by the application of DNase. The column is washed to remove any further contaminants, and then the RNA is eluted from the filter using water.

mRNA Structure

In eukaryotic cells, mRNA has a unique feature that allows researchers to either purify it away from the rest of the RNA or to direct enzymes to it specifically while avoiding the other RNA species. This feature is the polyA tail. The polyA tail is a long stretch of adenine nucleotides found at the 3′ end of mRNA, which is added post-transcriptionally. Such stretches of adenine nucleotides do not typically occur naturally in genes or other RNA species. The polyA tail will hybridize to an artificially generated oligonucleotide made up of a series of deoxythymine nucleotides (oligo-dT). If the oligo-dT is coupled to a support matrix (e.g., beads) the mRNA can be pulled out of solution thereby purifying it away from the rest of the total RNA. While some researchers prefer to include this step in their process, it is generally not a requirement for microarray analysis. Rather than purify the mRNA, the oligo-dT can be used as a primer for creating an enzymatically labeled complement of the mRNA.

Labeling

In order to render the RNA visible to a detection system, it is necessary to label it in some manner. While some laboratories choose a direct methodology of chemically labeling the mRNA itself, it is most common to work via a cDNA or cRNA intermediate that is labeled enzymatically.

The simplest methodology involves creating labeled cDNA. In this technique, the RNA is reverse-transcribed (DNA is made from an RNA template) by an enzyme named reverse transcriptase (RT) (for sample protocols, see Ref. 4). Reverse transcriptase requires a small oligonuclotide primer that binds to the RNA creating a short double-stranded region (an RNA:DNA hybrid). In order to ensure that the RT enzyme reads only the mRNA, the polyA tail of mRNA is exploited by using a primer made of a stretch of several (usually 20–25) thymine residues. The resultant DNA is the complement of the RNA and it is thus referred to as complementary DNA (cDNA). The RT reaction requires that free nucleotides (each of A, C, G, and T) are present to create the DNA. If one of these nucleotides is chemically modified with some detectable molecule (such as a fluorophore), then it will be incorporated into the cDNA strand, and that cDNA will be detectable with a fluorescent reader. Alternatively, it is possible to use a reactive molecule (such as amino-allyl) in place of a fluorescent molecule. After incorporation into the DNA, the DNA is then coupled to a reactive form of a fluorophore (usually a reactive ester). This latter implementation of the method has an advantage in that the amino-allyl modifier is a much smaller chemical group that is incorporated much more efficiently into DNA than a bulky fluorescent moiety.

Often the amount of RNA available is limiting and cannot be detected by standard means. In this case, it is generally necessary to amplify the amount of material present. A typical microarray experiment usually requires 5–10 μg of total RNA in order to be able to obtain useful data. When researchers are working with diminishingly small samples, such as from a needle biopsy or a fine needle aspirate, it is often not possible to obtain this amount of total RNA. To overcome this limitation, various amplification strategies have been adopted. The most popular method of amplification is based on the protocols of Dr. James Eberwine from the University of Pennsylvania (5). In this technique, RNA is converted into cDNA using the same method described above with two key differences: (1) there is no labeled nucleotide incorporated and (2) the oligo-dT primer has another short sequence of DNA appended to it that represents a T7 promoter region. The T7 promoter is a bacteriophage-derived sequence that initiates transcription by T7 polymerase. After the cDNA is created, a second strand is generated creating a double-stranded artificial gene with a T7 promoter on one end. This artificial gene is then transcribed by the addition of T7 polymerase, which is allowed to make numerous transcripts of the gene. The transcripts that are obtained can either be labeled directly, or they in turn can be turned into labeled cDNA using standard methodologies described above. The resultant RNA is now actually the opposite sequence of the original mRNA, so it is said to be cRNA (complementary RNA).

The Affymetrix GeneChips utilize an amplification system based on T7 transcription as described above. During the production of cRNA, biotin modified nucleotides are incorporated. Posthybridization (see the section on Hybridization) the arrays are stained with a streptavidin bound fluorophore. Streptavidin is a protein that specifically and tightly binds to biotin molecules, allowing the fluorophore to be attached to the cRNA.

A clean-up step is required to remove any free, unbound detection molecules. This step helps to ensure that background signal is kept to a minimum. There are two main methods by which such purification is performed, one is based on standard nucleic acid purification systems, similar to the RNA isolation method described earlier, and the other is based on size exclusion. For the first method, a nucleic acid purification column is utilized. The cRNA or cDNA binds to the silica filter, but the less charged free nucleotides flow through. After a series of washes, the cRNA or cDNA is eluted from the column. The second methodology utilizes a membrane filter (usually incorporated into a column) that has a defined pore size. The large cRNA and cDNA molecules are retained on the membrane; where as the small free nucleotides flow through. The column is then inverted and the cDNA or cRNA is then eluted off the column by flowing wash buffer in the opposite direction. This purified labeled material is then ready for hybridization to the array.

Hybridization

Microarray technology relies on the natural ability of single-stranded nucleic acids to find and specifically bind complementary sequences. Purified labeled material is exposed to the spotted microarray and the pool of labeled material "self-assembles" onto the array, with each individual nucleic acid (cDNA or cRNA) species hybridizing to a specific spot on the array containing its complement. The specificity of this interaction needs to be controlled, as there may be several similar and related sequences present on the array. The control of hybridization specificity is accomplished through the adjustment of the hybridization stringency. Highly stringent conditions promote exact matches where as low stringency will allow some related, but nonexact matches to occur. In a microarray experiment, stringency is typically controlled by two factors: the concentration of salt in the hybridization solution and the temperature at which hybridization is allowed to occur.

High salt concentrations tend to lead to lower stringency of hybridization. Both strands of nucleic acid involved in the hybridization event contain a net negative charge. As such, there is a small repulsion between these two strands, which needs to be overcome to bring the labeled nucleic acid into proximity of the arrayed probe. The salt ions cluster around the nucleic acid strands creating a mask and shielding the electrostatic forces. Higher salt concentrations have a greater masking effect, thus allowing hybridization to occur more easily. If salt concentrations are high enough, the repulsion effects are completely masked and even strands of DNA that have low degrees of homology may bind to one another.

Temperature is another important factor. Every double-stranded nucleotide has a specific temperature at which the two strands will "melt" or separate. The temperature at which exactly 50% of a population of pure double-stranded material separates is termed the melting temperature (T_m). The T_m of a nucleic acid is controlled partially by the length of the strand and partially by the percentage of G and C residues (termed the GC content). The G and C residues bind to one another as a Watson–Crick base pair. This pairing interaction is the result of three hydrogen bonds forming. The other potential base pair in a DNA hybrid, A:T, only has two such hydrogen bonds and thus the greater the GC content of the nucleotide, the more stable the hybrid. At very low temperatures, nonstandard Watson–Crick base pair interactions can also occur causing noncomplementary sequences or sequences that are $<100\%$ matched to form hybrids. It is necessary therefore to find a temperature that will prevent or melt nonspecific hybrids, but allow the specific interactions to occur. For a microarray, this presents a challenge as there are thousands of specific interactions that must be accommodated. In the case of oligonucleotide arrays, the design of the oligonucleotides to be spotted takes this issue into account and probes are designed that tend to fall within a narrow window of potential melting temperatures. cDNA arrays are more difficult because the sequences spotted vary greatly in both GC content and length. In such cases, it is often true that conditions that represent somewhat of a "compromise" are necessary.

Hybridization kinetics can generally be modeled as shown in Eq. 1(6). The change in the amount of hybridization product LS over time is a function of the decrease in the concentration of labeled target L and free spotted DNA S over time. To simplify the equation, the rate of hybridization is equal to some rate constant k multiplied by the product of the concentrations of L and S. Thus hybridization rate is a direct function of the concentrations of the labeled target molecule and the DNA probe in the spot.

$$\frac{d[\mathrm{LS}]}{d\mathrm{T}} = -\frac{d[\mathrm{L}]}{d\mathrm{T}} - \frac{d[\mathrm{S}]}{dT} = \frac{d[\mathrm{L-S}]}{dT} = k\,[\mathrm{L}][\mathrm{S}] \tag{1}$$

In the case of an oligonucleotide microarray, it is often the case that the number of spotted DNA molecules is in great excess to the number of target molecules. As such, the concentration of the spotted DNA probe remains fairly constant and can be considered part of the constant k. Thus the equation for hybridization can be simplified as shown in Eq. 2 (6), where the rate of hybridization is typically driven by the concentration of the labeled target molecules alone.

$$\frac{d[\mathrm{LS}]}{d\mathrm{T}} = k'[\mathrm{L}] \tag{2}$$

In the case of two color oligonucleotide arrays, the two labeled samples compete for hybridization to the probe that remains in excess and thus hybridization is simply a reflection of the concentrations of each of the two labeled targets $\mathrm{L_1}$ and $\mathrm{L_2}$ [Eq. 3(6)].

$$\frac{d[\mathrm{L_1S}]}{d[\mathrm{L_2S}]} = \frac{k'_1[\mathrm{L_1}]}{k'_2[\mathrm{L_2}]} \tag{3}$$

The situation becomes somewhat more complex when the probe molecules are not in excess of the target molecules. This is often the case with cDNA arrays. In these cases, the concentration of the spotted probe does change significantly as hybridization occurs and thus each of the labeled targets $\mathrm{L_1}$ and $\mathrm{L_2}$ hybridize in a manner described by Eqs. 4 and 5 (7).

$$\frac{d[\mathrm{L_1S}]}{d\mathrm{T}} = k_1[\mathrm{S}][\mathrm{L_1}] = k_1([\mathrm{S^0}] - [\mathrm{L_1S}] - [\mathrm{L_2S}])([\mathrm{L_1^0}] - [\mathrm{L_1S}]) \tag{4}$$

$$\frac{d[\mathrm{L_2S}]}{d\mathrm{T}} = k_2[\mathrm{S}][\mathrm{L_2}] = k_2([\mathrm{S^0}] - [\mathrm{L_2S}] - [\mathrm{L_1S}])([\mathrm{L_2^0}] - [\mathrm{L_2S}]) \tag{5}$$

In such a case, the rate of hybridization is affected by the change in the concentrations of the spotted probe from the initial concentration $\mathrm{S^0}$, where $\mathrm{S^0}$ changes as the probe molecules are bound by either $\mathrm{L_1}$ and $\mathrm{L_2}$.

When looking at differential hybridization between the two targets, we can represent the kinetics as shown in Eq. 6 (7).

$$\frac{d[\mathrm{L_1S}]}{d[\mathrm{L_2S}]} = \frac{k_1([\mathrm{L_1^0}] - [\mathrm{L_1}S])}{k_2([\mathrm{L_2^0}] - [\mathrm{L_2}S])} \tag{6}$$

If one is to assume that the two fluorescent molecules used in a two-color experiment behave similarly, and that the rate of hybridization of the two labeled targets is the same, we can say $k_1 = k_2$. It has been demonstrated that under ideal conditions and when the hybridization reaction is allowed to continue to equilibrium that the ratio of the concentrations of each possible hybrid $\mathrm{L_1S}$ and $\mathrm{L_2S}$ is equivalent to the ratio of the original concentrations of the two targets $\mathrm{L_1}$ and $\mathrm{L_2}$ [Eq. 7 (7)]. This point is important because it is the basis for microarrays to work, assuming that the ratios read from the scans during data analysis are reflective of an actual biological condition.

$$\frac{[\mathrm{L_1S}]}{[\mathrm{L_2S}]} = \frac{[\mathrm{L_1^0}]}{[\mathrm{L_2^0}]} \tag{7}$$

The goal of microarray hybridization is to produce a result for which the signal obtained from specific hybridization is very strong when compared to any background signal that may be obtained by a nonspecific adsorption of labeled material to the substrate, or nonspecific binding to spotted elements. To reach this goal, it is common to use certain nonspecific blocking reagents in the hybridization solution. Frequently, nucleic acids from sources known not to contain any sequences that will interfere with specific hybridization are used. For example, in a hybridization of a human sample to an array, one might use yeast tRNA and salmon sperm RNA as competitors to bind any regions of the substrate or probes that have a generic nucleic acid binding capacity. These nucleic acids are nonlabeled and will therefore not contribute any signal when the array is scanned.

Washing

Unlike traditional northern blots, the majority of the stringency of a microarray assay is accomplished at the hybridization step. The washing step of a microarray experiment is a critical operation, but is important more as a means to remove unbound material in order to reduce background signal than it is to control the specificity of the signal obtained.

Wash buffers generally contain two components: a salt solution and a detergent. The salt solution, frequently sodium chloride sodium citrate (SSC), is set to a concentration that supports the maintenance of the hybridized molecules. This concentration most frequently falls in the 1× to 2× concentration range with some labs using as low of a concentration as 0.1× (1× SSC contains 0.15 *M* NaCl and 0.015 *M* Na-citrate).

The detergents used in wash buffers help to remove the unbound fluorescent molecules that would normally stick to the surface of the slide. The detergent acts as a surfactant and helps to isolate and remove the unbound fluorescent material. Typically, an anionic detergent such as sodium dodecyl sulfate (SDS) is used for this purpose.

The temperature for the washes varies depending on the stringency of the wash solution being used. As with hybridization, the combination of temperature and salt concentration determines the overall stringency of the washes.

After washing the microarrays, it is generally necessary to perform a rinse. The rinse is typically a solution similar to the wash solutions without the detergent. If detergent remains on the slide after drying, the solution may fluoresce particularly if the labeled material has been trapped in detergent micelles.

Scanning

It is necessary to use an imaging device to detect the fluorescent labels present on the hybridized microarray. In general, the imaging device must contain an excitation light source, an emission filter, and a light gathering device.

During scanning, the labeled material, be it fluorescent or some other form of detectable molecule, is imaged and the resultant data is converted to a digital image. The optimal resolution at which the image is scanned is dependent on the size of the features and on their interspot spacing. A general rule of thumb is that the resolution of the image should be such that the pixels represent one-tenth of the diameter of the spot. For spotted arrays, for example, the features tend to be on the order of 100 μm in diameter and thus 10 μm resolution is frequently used. Affymetrix's technology, however, can generate features that are 11 μm square; in this case, a much higher resolution of down to 1 μm is required.

Most commonly, the image that is generated is a 16-bit grayscale TIFF (Tagged Image File Format) image (Fig. 2). The 16-bit depth of the image provides a total of 65,536 gray levels providing a possibility of more than five orders of magnitude range. The TIFF format is important because it is a universally accepted format that is LOSSLESS; that is, even with compression, this format retains all image information. The images can then be imported into the appropriate image quantification software.

Image Quantification

After scanning, it is necessary to extract data from the images. Image quantification generally starts with segmentation. Segmentation is the process by which pixels that represent the signal are isolated from those that represent background. During segmentation, the discrete areas of the image that represent the spotted DNA material are identified and digitally isolated from the remainder of the image. The intensities of all of the pixels in the individual spot are averaged to determine the overall spot intensity. This spot intensity is proportional to the amount of material hybridized to that region, with higher intensities resulting from increased numbers of hybridized molecules. Each spot, for each channel (in the case of two color microarrays) is quantified, and the resultant data are tabulated. Other data may also be extracted at this stage. It is common to also obtain intensity data for the area outside of the individual spots. This value represents the background of the image and indicates the amount of signal that would have been obtained regardless of a specific hybridization event. It is common, however, not universal, to subtract the background values from the signal intensities of the spots.

There are several means by which segmentation can be carried out. In the most basic setup, a fixed shape (usually a circle) is placed over each spot. The entire complement of pixels lying within the circle is used to determine the average intensity. Pixels lying outside of one of these circles are deemed to be background signal. More advanced segmentation algorithms attempt to account for the fact that most of the spotted features on a microarray are in fact not perfectly uniform. Spots may deviate from a true circular shape, or may have regions within the circle in which DNA was not attached (creating a spot that is reminiscent of a doughnut). In addition, it is not uncommon for each of the spots to have some degree of variance in their diameter. The more advanced methods utilize various algorithms and statistics to determine which pixels actually represent signal and which are more representative of background.

Image quantitation software then processes the entire image and produces a table of results that represents the signal, and the background for each feature on the array. These packages may also export various other data, which can be used in quality control analysis such as standard deviations, coefficients of variance, circularity, or uniformity of the spot, and so on. This data table can then be processed as part of the data analysis.

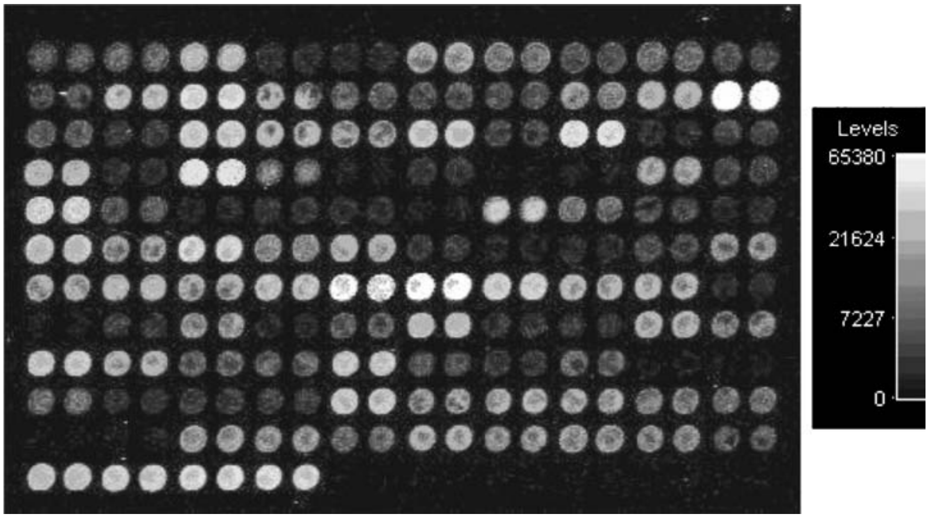


Figure 2. Arrays imaged on a microarray scanner are presented as 16-bit grayscale TIFF images. The picture shown represents a small subsection of a larger array. Each spot is 100 μm in diameter and the spot-to-spot spacing is 200 μm in this image. The image was scanned at 10-μm resolution.

Data Analysis

An exhaustive description of the process of DNA microarray data analysis is far beyond the scope of this article (for an excellent review see Ref. 8). The exact process followed depends greatly on the experimental design and the question being addressed. There are, however, some basic principles that tend to be fairly common in dealing with microarray data: statistical analysis of data, supervised and/or nonsupervised data mining, data visualization, and validation are all key components.

Statistical analysis of microarray data comes into play in two main areas. The first is to determine which spots are reliable and provide sufficient data. Spots that have a high degree of variance across replicates, for example, are likely not able to provide reliable data. These hypervariable genes or signals need to be filtered from the data so as to not skew the results of data mining. Statistics may also play a role in supervised data analysis.

There are two major categories of data mining: supervised and nonsupervised. Supervised data mining utilizes algorithms in which the user imparts restrictions on how the data is grouped. For example, in an experiment where a cohort of patients was tested in which one group was healthy and the other group was afflicted with a particular disease, one would indicate to the algorithm which arrays were from the healthy patients and which were from the patients with disease. The algorithm then tests the data to find genes that are markers for the diseases. Specifically, each gene is tested to see if the expression levels for that gene are statistically significantly different in each of the two patient groups. The goal is to find a series of genes that can act as markers that are diagnostic of the disease.

In nonsupervised clustering, the algorithm is not given any indication as to how the individual samples are related. In true nonsupervised clustering, the algorithm is not even told how many groups exist. The data are analyzed and the samples are grouped based on similarity metrics. The classical methods of nonsupervised clustering include hierarchical clustering and principal components analysis (PCA). The algorithms generally display the data via some visualization pattern such as the canonical "plaid" expression patterns seen from hierarchical clustering. The researcher then overlays the grouping information onto the patterns provided to see if the individual groups naturally separate from one another. In other cases, this methodology may be being used to determine how many groups there truly are, as the researcher may not have this information a priori. In such cases, the groups can then be further examined to see if there are differences in treatment response, survival, or any other characteristic desired. Generally, after this technique is performed one will attempt to look for clusters of genes in the patterns that distinguish between the different groups and again use these genes as markers.

Regardless of the methodology utilized, it is extremely important to validate the data. Cross-validation strategies are various, but in their most basic form, one obtains a cohort of patients to profile. A subset of this cohort is used to look for potential markers. Once the markers have been identified, the remaining patients are tested and only the identified markers are used to try and group the patients. If the markers are able to stratify the patients into their appropriate groups, then the markers are considered to be viable and may provide beneficial diagnostic ability. On occasion, however, the validation set is not properly grouped. In such cases, the markers are only useful for the narrow set of patients used in the initial tests and more testing is required to find a viable set of markers.

FABRICATION OF MICROARRAYS

There are two main methodologies for manufacturing microarrays, which differ in the means by which the probe material spotted onto the arrays is prepared. In one methodology, the DNA to be spotted is generated *in situ* using either standard or modified phosphoramidite chemistry. (Phospohoramidites are reactive forms of each of the nucleotides that make up DNA. Phosphoramidite chemistry is a well-defined process by which moderate length stretches of DNA can be created with any specific sequence.) This method is used by Affymetrix and Agilent, the two largest commercial suppliers of microarrays, although both groups use a different approach to the *in situ* synthesis.

Other groups use *ex situ* synthesis, whereby the DNA material is either prepared as PCR products (cDNA) or oligonucleotides manufactured using standard phosphoramidite synthesis. Once this material is prepared it is spotted onto the array substrate using either contact or noncontact printing methodologies. Amersham (now GE Healthcare) and Applied Biosystems use this methodology to make microarrays as do almost all of the "homebrew" laboratories that make microarrays in house.

Fabrication of DNA Arrays *In Situ*

There are two main approaches to the generation of microarrays by *in situ* synthesis of DNA: photolithography and inkjetting. Affymetrix, the industry leader uses a proprietary photolithography process to mask off areas of the array, protecting some areas, and leaving others available for the DNA synthesis reaction to occur (1). This is a multistep process requiring several masks per array to be made. Each synthesis reaction is performed sequentially. For each nucleotide position, there are four possible masks (one for each of A, G, C, and T). Thus, an array comprised of 25-mer oligonucleotides would require ~100 masks to complete the process (typically ~70 are required for an array due to the sequences used). Affymetrix uses a modified phosphoramidite chemistry for synthesis of the oligonucleotide chains; whereas standard phospohoramidite chemistry uses acid labile protection groups, the Affymetrix technology utilizes groups that can be removed by ultraviolet (UV) light. The Affymetrix technology allows for extremely high density arrays of hundreds of thousands of features to be prepared on very small substrates of $<1\ \mathrm{cm}^2$.

Other groups have developed technologies that allow them to get around the need for multiple masks to be made for each array design. The pioneer in this area was Nimblegen, who uses digital light processor (DLP) micromirrors to create the masks (9). Each of these DLP units (used typically in AV projectors and large screen televisions) comprises thousands of tiny (10 μm^2) micromirrors. The micromirrors can be individually addressed and the angle of the mirrors changed to allow light to pass through. In the "open state", the micromirror directs light onto the surface of the microarray, allowing DNA synthesis to occur. In the "closed state", the micromirror reflects light away from the surface, disallowing DNA synthesis. A computer controls the mirrors and thus each DLP unit has a near infinite number of combinations that can each be controlled, and as such, a single unit can create any pattern desired on the array. Nimblegen uses the same chemistry as Affymetrix, using light activated deprotection of the phosphoramidites. A somewhat newer entry into this area is Xeotron (now part of Invitrogen). Xeotron also uses micromirror DLPs to address the masks, however, they have also incorporated small microfluidic channels on their chips. Each feature is placed in a microscopic well on the chip. Rather than using the modified phosphoramidite chemistry of Affymetrix and Nimblegen, Xeotron uses standard chemistry, but has instead employed a caged acid that can be freed by light (10,11). As such, the acid that controls deprotection of the nascent oligonucleotide can be directed to specific locations by light. The Nimblegen and Xeotron technologies have the advantage of being highly amenable to custom array generation, however the Affymetrix technology is particularly well suited to mass production of a standard array. Each of these approaches has found customers in the marketplace.

A third approach to *in situ* synthesis of the oligonucleotides involves ink-jet spotting. Agilent uses this technology (developed by Rosetta Inpharmatics) in which each of the reactive phosphoramidites (A, G, C, and T) are loaded in to a separate "ink-cartridge" to allow for control of which nucleotide is added to each spot during the synthesis stage (12,13). This methodology eliminates the need for masks, but does require very high precision robotics as the print head must return to the same spot many times, within micron accuracy, during the course of synthesis. This technology draws from the strength of each of the others mentioned in that it is relatively easy to customize the design of arrays, and yet, mass production of arrays is possible using a large robotic system.

Fabrication of DNA Arrays *Ex Situ*

Some of the commercial vendors and nearly all of the "homebrew" microarray centers utilize and approach of spotting DNA that was prepared *ex situ*. In the case of cDNA arrays, the spotted material is prepared by polymerase chain reaction (PCR), whereas oligonucleotide arrays are generated using oligos created via high throughput oligo synthesis. The DNA material is purified and placed into a specific spotting buffer that is compatible with the substrates being used.

The DNA is typically aliquoted out into multiwell plates (96, 384, or 1536 wells /plate) to facilitate transfer by the arraying robot. The buffer that the DNA is placed in has several functions. First, the buffer stabilizes the DNA to prevent it from degradation. Second, the buffer must provide an appropriate surface tension to ensure that the spots that are placed on the substrate are of a controllable size and uniform in shape. Of similar importance, however, is that the buffer must provide conditions that are compatible with the attachment chemistry that is going to be utilized.

The DNA may either be coupled to the slide through rather simple electrostatic interactions or via a specific coupling reaction. Electrostatic interactions are mediated by using a uniform positively charged substrate that attracts the negatively charged DNA. Often the substrates used are silylated to provide reactive amine groups on the surface. Alternatively, one may coat the slides with a chemical such as poly-L-lysine, which simply adsorbs onto the substrate and provides a net positive charge. This type of interaction is mass based. As such, there is a maximum mass of DNA that can bind to any one spot on the substrate. Longer DNAs will be represented by fewer copies than shorter DNAs. To overcome this, it is possible to use more specific interactions by using modifiers on the DNA that will react with certain groups on the slide. The two most common such modalities involve aldehyde or epoxide chemistry. In this method, the DNA is modified with a primary amine group. The substrate has reactive aldehydes or epoxides that will react specifically with the primary amine to form a covalent bond (Fig. 3). This type of interaction is molarity based, and as such, with the exception of steric effects, the number of DNAs that bind per spot is relatively equivalent regardless of length.

EQUIPMENT

The manufacture of microarrays, and their subsequent use requires some very specialized equipment. Generally, a facility that produces microarrays will require some advanced robotics for fabrication. A laboratory that uses arrays will require scanning devices to read the arrays. Due to the relatively high costs of these pieces of equipment it is common for many people to rely on core facilities for some or all of the process.

Arraying Robots

Ex situ prepared DNAs are spotted onto the microarray substrates via robotics (Fig. 4). Robotics are required to accurately position the printing devices over the slides to create the arrays. The majority of systems utilize pins and direct contact to deposit the DNA material. In this system, a printhead with several spotting pins in a defined arrangement is used to dip into the multiwell plates and pick up the material to be spotted. The typical operation sequence of an arrayer robot may include:

1. Dipping the printing applicators (pins) into a source plate to pick up DNA samples. Each applicator picks up a separate DNA sample from an individual well in the plate. Typically 32–48 pins are used at one time.
2. Movement to a blot-station to preprint from the pins. This step removes excess solution from the pins to ensure that the spots that are printed onto the arrays

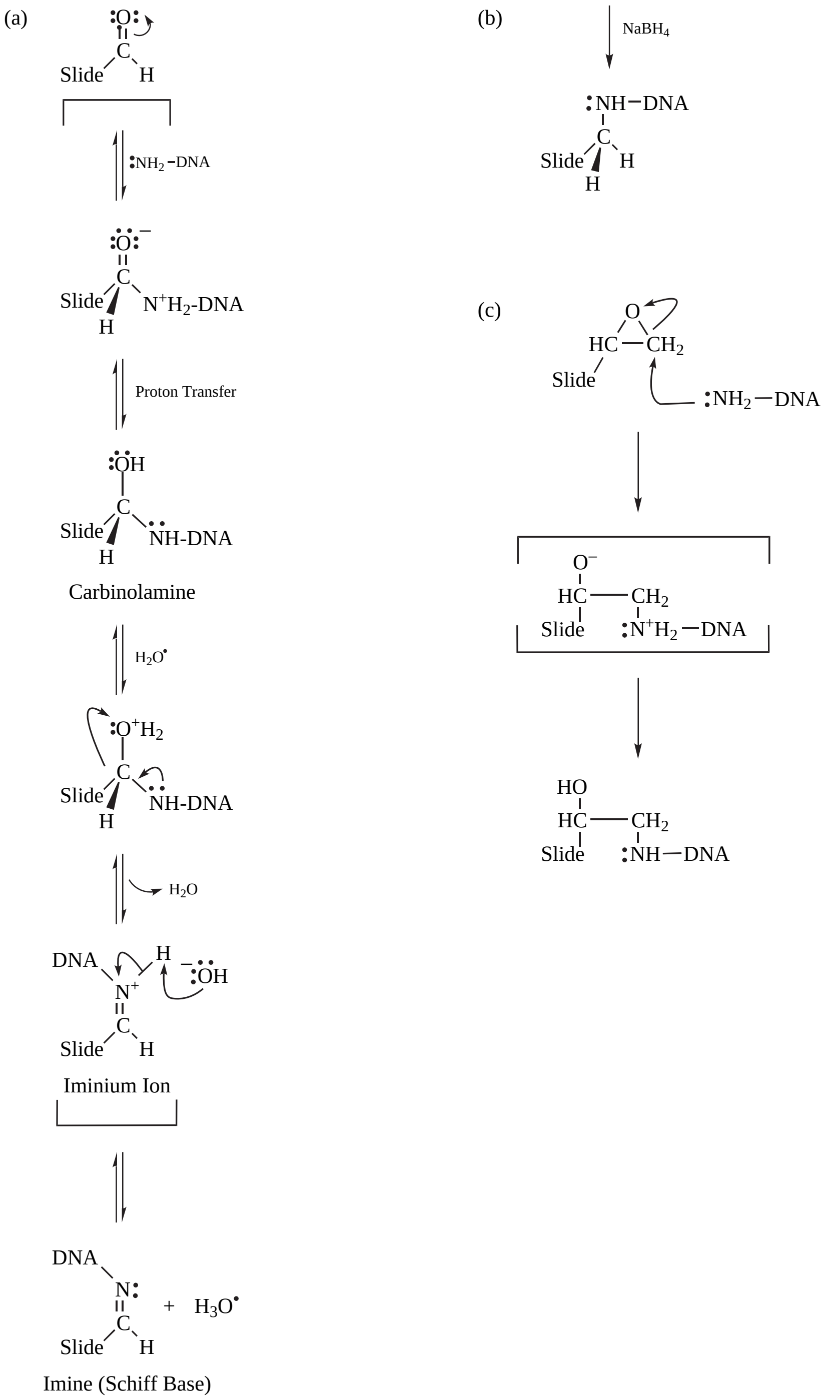


Figure 3. Covalent attachment of amino-modified DNAs to aledhyde (a) or epoxide (b) slides is possible. An amino-modified DNA reacts with an aldehyde surface by a Schiff's base reaction. The resultant Schiff base must be reduced with an agent such as sodium borohydride ($NaBH_4$) to prevent reversal of the reaction.

are uniform in size and do not run into one another causing contamination.

3. Movement to the slide platform. The print head then moves over the slide platform taking position over the first slide.
4. Printing onto the arrays. The print head moves down bringing the pins in contact with the slide. The DNA solution held in the pins by capillary action is spotted onto the slide. The printhead then moves to the next slide position and again spots onto the slide. This process is repeated until all of the slides on the platform have been printed.
5. Washing the pins. The print head then moves the pins to a wash station. Although there are many configurations possible, the basic principle is to use water or some other solution to remove the excess liquid from the pins and then to dry the pins (under vacuum or stream of air). This process may be repeated several times to make sure there is no carryover.

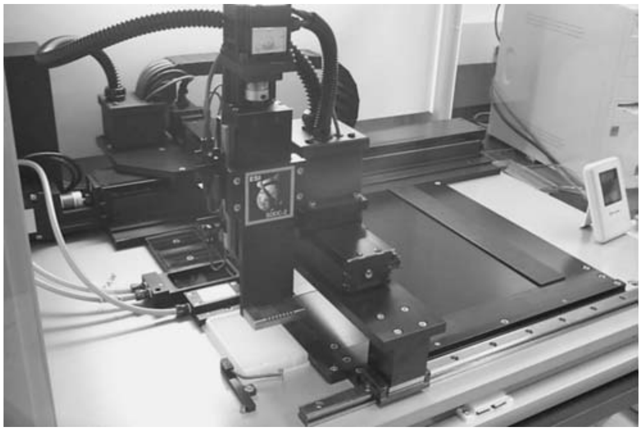


Figure 4. A microarraying robot. The robotic arrayer prints DNA onto glass slides with very high precision. Robots such as this have extremely high accuracy, on the order of 10 μm or less.

6. Loading the next sample. The print head returns to the source plate to pick up the next set of samples.

In a typical high throughput system, such as those offered by Bio-Rad, BioRobotics, GeneMachines, Genetix, and Telechem International, 48 pins are used at one time. The entire operation sequence described above may take 3–4 min to complete for 100 arrays. Often arrays may contain 20,000–40,000 spots. As such, a typical print run may require 600 or more cycles through the operation sequence, which can take as long as 30 h or more to complete.

Hybridization and Fluidics Stations

Certain array platforms require that a specific hybridization and/or fluidics station be utilized. In the case of spotted arrays (home-brew in particular), this is usually an option and often a case of personal preference. In these cases, a hybridization station may be utilized to improve mixing of the hybridization solution over the array. The rate of diffusion of a labeled nucleic acid in solution is actually very low, and as such, some researchers prefer to use an automated station that performs mixing of the solution.

In the case of Affymetrix GeneChip technology, a specific hybridization and fluidics station are required. The hybridization station is simply a rotating incubator in which the chips are placed. A bubble that is introduced into the sealed array cartridge moves around during rotation creating a mixing effect. The fluidics station is a more advanced system that is required to introduce the various labeling components and wash solutions required. This station allows the user to keep the cartridge sealed without having to attempt to pipette solutions in and out.

Scanners

While some microarray imagers such as the Perkin Elmer ScanArray and GeneFocus DNAScope are confocal scanners, this is not a strict requirement. Confocal imaging serves to eliminate extraneous signals, but reduces the light gathering ability of the device. There are >10,000 commercial microarray scanners in the field capable of reading standard glass microarrays. The leading scanner makers include Agilent, Axon, Bio-Rad, GeneFocus, PerkinElmer, and other vendors. The laser scanner uses one or more lasers with wavelengths appropriate to the fluorophores being used. The most commonly used fluorophores for microarrays are cyanine 3 and cyanine 5 (or fluors with equivalent spectra). Cyanine 3 has an absorbance maximum of 550 nm and emission maximum of 570 nm. There are 2 main lasers used in scanners to excite this fluorphore: “Gre-Ne” (green neon) gas lasers and Nd:YAG (neodymium doped yttrium aluminum garnet) frequency doubled solid-state diode lasers. Cyanine 5 has an absorbance maximum of 650 nm and an emission maximum of 670 nm. There are two main lasers used in scanners to excite this fluorophore: standard He–Ne gas lasers and red diode lasers. Table 1 shows some of the characteristics of these two dyes, along with two other popular dyes, Alexa 555 and Alexa 647, which have spectra that are very similar to those of Cy3 and Cy5 respectively (Fig. 5).

Cyanine 3 and 5 have some important features that make these dyes particularly suitable for use in microarray analysis. The spectra of these dyes have little over lap and can generally be separated from one another with little to no cross-talk. In addition, these fluors have a somewhat unique property in that they are brighter when dry than when wet. Most fluorophores have the opposite behavior, which is impractical for microarrays because the scanners generally cannot handle wet preparations.

The other major class of microarray imager is a CCD (charge coupled device) based system. In general, these imagers use a white light source to excite the fluorophores. The fluorescent light that is emitted is captured by the CCD and converted into a digital image. Rather than scanning the slide, a CCD based imager tiles together several sections of the slide to create an image of the entire surface. This tiling can create a stitching effect whereby the “seams” of the images may not be completely smooth.

Table 1. Key Characteristics of the Most Commonly Used Fluorophores for Microarray Analysis

Fluorophore	Excitation Max, nm	Emission Max, nm	Molar Extinction Coefficient	Molecular Weight
Cy3	550	570	150,000	766
Cy5	649	670	250,000	792
Alexa555	555	565	150,000	1,250
Alexa647	650	668	239,000	1,250
Phycoerytherin	566	575	19,600,000	240,000

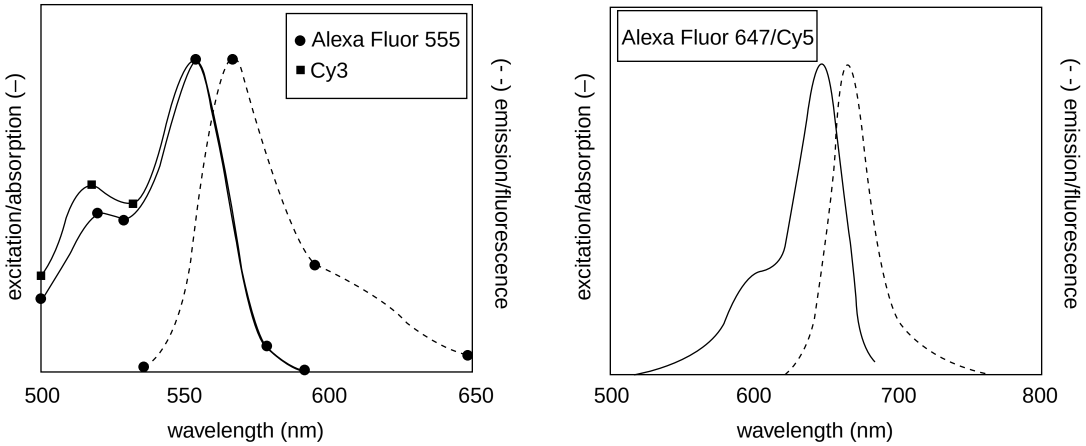


Figure 5. Representative spectra of the fluors commonly used in spotted microarray experiments. Alexa Fluor 555 and Cy3 are excited by green wavelengths of light whereas Alexa Fluor 647 and Cy5 are excited by red wavelengths of light. One green excited and one red excited fluor may be used at the same time as there is little overlap in their excitation spectra.

This problem can be overcome with advanced lighting systems and software.

Affymetrix arrays use a different labeling chemistry for detection relying on the naturally occurring fluorescent protein phycoerytherin. Phycoerythrin is a naturally occurring pigment protein from light harvesting algae that absorbs strongly at 566 nm and has an emission peak at 575 nm. It is a very bright fluorophore having a molar extinction coefficient that is 80 times as high as the standard Cy3 and Cy5 molecules. The limitation of this molecule is that it is also 200 times larger, making the number of molecules that can be incorporated per sequence much less. As such, this molecule can only be applied to the DNA posthybridization for fear that it would create steric interference.

MICROARRAYS AS MEDICAL DEVICES

To date, microarrays have mostly found use in basic research applications, and have yet to make a strong impact on the diagnostic market. [During the preparation of this text, Roche received FDA clearance for the first ever array based diagnostic chip. The AmpliChip CYP450 based on the Affymetrix platform was approved in January of 2005 (see http://www.roche.com/med-cor-2005-01-12).] Microarrays have indeed been used to study many diseases including various cancers, cardiovascular disease, inflammatory disease, psychiatric disorders and infectious disease. This basic research will ultimately lead to the identification of potential therapeutic markers for drugs of for diagnostics. The potential of microarrays extends beyond target discovery, however, and will eventually impact on the way that medical care is performed.

Target Discovery

The use of microarrays in basic research laboratories has often focused on target discovery. In these applications, microarrays are used to profile a particular disease where disease tissues are compared to healthy tissues either from the same patient or from a separate test population. In such experiments, the goal is to find genes that are differentially regulated (either up or down) in the disease state compared to a healthy tissue. Such genes are thought to be involved in the disease state or in the cellular response to the disease. As such, these genes are potential diagnostic markers and may also represent drug targets.

Drug/Lead Discovery

Microarrays can also be used once the target has been identified. It is possible to use microarrays to screen potential therapeutic compounds, for example, to determine which candidates reverse the pattern of gene expression that is indicative of disease. Microarrays have been even more effective in looking at toxicity of lead compounds. One of the leading contributors to failure of a pharmaceutical compound is toxic or off target events. Microarrays have proven useful in screening for the up-regulation in toxicity related genes. In addition, it is possible to determine if the compound creates other effects that while not toxic *per se* could cause undesirable side effects from nonspecific interactions. Often toxicity models are tested in model organisms such as rats or dogs. Several toxicity specific arrays have been developed that allow for profiling of genes in these model systems rather than human cells.

Diagnostics and Prognostics

One of the more promising areas for microarrays to have direct impact as a medical device is in the area of diagnostics and prognostics. As mentioned under target discovery, basic research has often strived to look for a panel of genes that can be used as a molecular fingerprint of a disease. There are numerous publications in which researchers have attempted to use molecular profiles to correlate to patient outcome, disease state, tumor type, or any of several other factors. DNA microarrays are particularly well suited to this type of analysis. Many complex diseases are multifactoral; rather than a single prognostic or diagnostic marker being present, it may be necessary to look at several genes at one time. Microarrays allow for identification of a panel of genes, which when looked at together may provide diagnostic or prognostic power. Although it has not become common practice yet, there are examples of microarrays being used to prescreen patients on the basis of a molecular profile (14).

Other attempts are being made at using microarrays to study infectious disease. Often times a patient may present with a set of symptoms that could be indicative of several different infectious agents. It is possible to prepare a microarray that would identify the agent as well as to subtype the bacterium or virus on the basis of pathogenicity. This particular application may prove very useful in identifying not only the infectious agent, but also the best course of treatment.

Pharmacogenomics and Theranostics

A concept that is gaining in popularity is pharmacogenomics or theranostics (15). Both of these terms refer to the idea of tailoring a patient's treatment or therapy on the basis of their genetic makeup. Many pharmaceuticals on the market have not known any potentially serious side effects in a subset of patients. In addition, there are typically at least some patients that are nonresponders to a particular treatment. These effects are often times the result of the patient's genetic make-up. Most of the work in this area has focused on genotyping: looking at certain variable regions of DNA and determining which variants are present in people who have negative reactions or in people who respond well to a treatment. It is hoped that in the near future it will be possible to screen a patient and determine which of a panel of drugs will be most beneficial. Perhaps even more important, it will be possible to prevent serious negative outcomes by avoiding treatment of a patient that will have a poor reaction to a drug. Theranostics also involves monitoring a patient through a course of treatment. It is possible that a patient can be screened during treatment to ensure that the therapy is working as expected. If a change occurs, the physician would be able to alter the therapy to ensure that the disease is treated in the most effective way possible.

SUMMARY

Microarrays provide a means to screen hundreds to thousands of biological analytes in parallel. These analytes can be DNA, RNA, or protein. DNA microarrays allow for rapid profiling of gene expression. While there are a few competing platforms that can be utilised, the basic principles are the same: RNA from a biological sample is extracted, labeled and applied to an array of DNA probes. Signals generated from the array indicate which genes are active and which are not. The ability to screen multiple tissues or patients make microarrays particularly well suited to uncovering the complex gene networks involved in disease. While typically used in basic research applications for target or marker discovery, the future will most likely see microarrays used in diagnostic applications and for tailoring medical treatment.

BIBLIOGRAPHY

1. Fodor SP, Rava RP, Huang XC, Pease AC, Holmes CP, Adams CL. Multiplexed biochemical assays with biological chips. Nature (London) 1993;364:555–556.
2. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995;270:467–470.
3. Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, Kampa D, Piccolboni A, Sementchenko V, Cheng J, Williams AJ, Wheeler R, Wong B, Drenkow J, Yamanaka M, Patel S, Brubaker S, Tammana H, Helt G, Struhl K, Gingeras TR. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding rnas. Cell 2004;116: 499–509.
4. Hegde P, Qi R, Abernathy K, Gay C, Dharap S, Gaspard R, Hughes JE, Snesrud E, Lee N, Quackenbush J. A concise guide to cdna microarray analysis. Biotechniques 2000;29: 548–550, 552–544, 556 passim.
5. Van Gelder RN, von Zastrow ME, Yool A, Dement WC, Barchas JD, Eberwine JH. Amplified rna synthesized from limited quantities of heterogeneous CDNA. Proc Natl Acad Sci USA 1990;87:1663–1667.
6. Schena M. Microarray analysis. Hoboken: John Wiley & Sons; 2003.
7. Wang Y, Wang X, Guo SW, Ghosh S. Conditions to ensure competitive hybridization in two-color microarray: A theoretical and experimental analysis. Biotechniques 2002;32: 1342–1346.
8. Quackenbush J. Computational analysis of microarray data. Nature Rev Genet 2001;2:418.
9. Singh-Gasson S, Green RD, Yue Y, Nelson C, Blattner F, Sussman MR, Cerrina F. Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. Nat Biotechnol 1999;17:974–978.
10. Gao X, LeProust E, Zhang H, Srivannavit O, Gulari E, Yu P, Nishiguchi C, Xiang Q, Zhou X. A flexible light-directed DNA chip synthesis gated by deprotection using solution photogenerated acids. Nucleic Acids Res 2001;29:4744–4750.
11. LeProust E, Pellois JP, Yu P, Zhang H, Gao X, Srivannavit O, Gulari E, Zhou X. Digital light-directed synthesis. A microarray platform that permits rapid reaction optimization on a combinatorial basis. J Comb Chem 2000;2:349–354.
12. Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, Lefkowitz SM, Ziman M, Schelter JM, Meyer MR, Kobayashi S, Davis C, Dai H, He YD, Stephaniants SB, Cavet G, Walker WL, West A, Coffey E, Shoemaker DD, Stoughton R, Blanchard AP, Friend SH, Linsley PS. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nat Biotechnol 2001;19:342–347.
13. Hughes TR, Shoemaker DD. DNA microarrays for expression profiling. Curr Opin Chem Biol 2001;5:21–25.
14. Schubert CM. Microarray to be used as routine clinical screen. Nat Med 2003;9:9.
15. Picard FJ, Bergeron MG. Rapid molecular theranostics in infectious diseases. Drug Discov Today 2002;7:1092–1101.

See also DNA SEQUENCE; MICROBIOREACTORS; POLYMERASE CHAIN REACTION.

MICROBIAL DETECTION SYSTEMS

Patricia L. Dolan
Jason C. Harper
Susan M. Brozik
Sandia National Laboratories
Albuquerque, New Mexico

INTRODUCTION

Infectious diseases accounted for 25–33% of the estimated 54 million deaths worldwide in 1988 (1), of which more than half are attributed to tuberculosis, malaria, chronic hepatitis B, diarrheal diseases, and human immunodeficiency virus/Acquired Immune Deficiency Syndrome HIV/AIDS. The incidence of more than 30 diseases identified since the mid-1970s continues to grow, which include HIV/AIDS, liver disease due to hepatitis C virus, cholera, tick-transmitted Lyme disease, foodborne illness caused by *E. coli* O157:H7 and *Cyclospora*, waterborne disease due to *Cryptosporidium*, and the hantavirus pulmonary syndrome. Additionally, the first known cases of human influenza caused by the avian influenza virus, H5N1, were identified in Hong Kong in 1997 (2).

Although death due to infectious diseases in the United States remains low relative to that of noninfectious diseases, their occurrence is increasing. In 2000, the Federation of American Scientists reported that infectious-disease-related death rates nearly doubled from 1980 to 170,000 annually (1). Many of these diseases, most recently the West Nile virus, were introduced from outside the U.S. borders by international travelers, immigrants, returning U.S. military personnel, or imported animals and foodstuffs. Still, the most dangerous infectious microbes reside within U.S. borders. Four million Americans are chronic carriers of the hepatitis C virus, a significant cause of liver cancer and cirrhosis. It is predicted that the death rate due to hepatitis C virus infection may surpass that of HIV/AIDS in the next five years. Influenza viruses are responsible for approximately 30,000 deaths annually. In addition, hospital-acquired infections are surging due to highly virulent and resistant pathogens such as *Staphylococcus aureus*.

The burden of identifying and treating infected individuals and controlling disease outbreaks generally lies with physicians, hospitals, and first responders. Table 1 contains important characteristics for several of the more common pathogenic microorganisms. As evidenced by this noninclusive table, a wide variety of microorganisms exists from which the specific diseasecausing microbe must be identified. In addition, the number of cells or particles that can provide an infectious dose is often extremely low. For example, the infectious dose of *E. coli* O157:H7 is as low as 10 cells (3), which poses a significant challenge to healthcare professionals and first responders who must quickly identify the infectious agent. Antimicrobial treatments that attempt to neutralize all possible infectious pathogens are often not possible or safe. Depending on the nature and severity of the infection, a delay of only a few hours in providing the proper therapy may lead to death.

Medical Microbiology

Medical microbiology is the discipline of science devoted to identifying microbial agents that are responsible for infectious disease and elucidating the mechanism of interaction between the microorganism and human host. Historically, microbiologists have used plating, microscopy, cell culture, and susceptibility tests to identify and study microorganisms. In the hospital and clinical diagnostic laboratory, these, methods are still widely used and will be briefly discussed in this article. The general procedure for isolation and identification of infectious and parasitic microbes is (1) specimen collection and streaking onto culture plates for production of isolated bacteria colonies, (2) staining and microscopic analysis, (3) cell culture in various media, and (4) antibiotic susceptibility testing.

Plating. Plating entails the streaking of a specimen onto a solid nutrient media-filled Petri dish and incubation at 35–37 °C. Under these conditions, a single bacterium divides and eventually produces a colony that is visible to the eye (Fig. 1). A visible colony generally contains more than 10^7 organisms. The colony morphology, color, time required for growth, appropriate media, and other growth conditions are used to characterize the microbe. The incubation time required for growth of a colony from a single cell is dependent on the growth rate of the microorganism. Fast-growing organisms such as *E. coli*, with a doubling time of 30 minutes, would require approximately 13 hours to produce a colony of 10^7 organisms. More typically, microorganisms require several days to a week to generate visible colonies. The plated specimen is often a complex solution such as blood, urine, feces, or sputum containing diverse native flora in addition to the infectious microbe. Plating serves as a method to isolate the infectious microbe as each

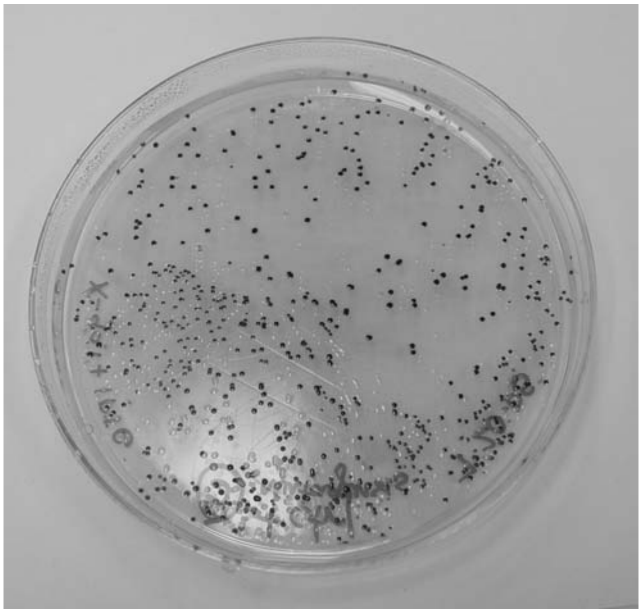


Figure 1. Colonies of *E. coli* grown on LB +ampicillin and X-gal +IPTG agar media in a Petri dish.

Table 1. Characteristics of common pathogenic organisms

Organism	Site(s)	Disease(s)	Incubation period	Mode of transmission	Infectious Dose
Bacteria, Gram positive					
Corynebacterium diphtheriae	upper respiratory track, skin	diphtheria	2–7 days	direct contact, droplet spread	unknown
Streptococcus pneumoniae	lower respiratory tract	pneumonia, meningitis	1–3 days	direct contact, droplet spread	unknown
Listeria monocytogenes	monocytes, leukocytes, blood, intestine, skin	listeriosis, meningoencephalitis	variable; 3–70 days	ingestion, direct contact, neonatal contact (birth), transplacental	unknown, <1000 cells
Bacillus anthracis	lower respiratory tract, skin lesions, intestine	anthrax	few hours −7 days	direct contact, airborne, ingestion	8,000–50,000 spores
Staphylococcus aureus	skin, osteomyelitis, blood, heart	boils, furuncles, abscesses, impetigo, osteomyelitis, sepsis, toxic shock syndrome	variable; 4–10 days	direct contact, ingestion, autoinfection, neonatal contact (birth)	$10^2 – 10^6$ cells
Streptococcus pyogenes	throat, skin, blood, middle ear	pharyngitis, septicemia, erysipelas, rheumatic fever, scarlet fever, otitis media, foodborne illness	1–3 days	direct contact, droplet spread, injection	<1000 cells
Bacteria, Gram negative					
Bordetella pertussis	oropharynx	whooping cough	6–20 days	direct contact, droplet spread, airborne	unknown
Escherichia coli(O157:H7)	large intestine	hemorrhagic colitis	2–8 days	ingestion	∼10 cells
Legionella spp.	lower respiratory tract	Legionellosis, Pontiac fever	2–10 days	airborne (non–communicable)	unknown
Neisseria gonorrhoeae	genitourinary tract, eye	gonorrhoea, pelvic inflammatory disease, septicemia, pharyngitis	2–7 days	direct contact (usually sexual), neonatal contact (birth)	unknown
Neisseria meningitidis	meninges	meningitis	2–10 days	direct contact, droplet spread	unknown
Salmonella spp.	small intestine	gastroenteritis	6–72 hours	ingestion	10–100,000 cells
Salmonella typhi	small intestine	typhoid fever	1–3 weeks	ingestion	1,000–100,000 cells
Shigella spp.	intestine	shigellosis (enteritis); Bacillary dysentery	1–7 days	ingestion	10–200 cells
Yersinia spp.	intestine	enterocolitis, acute mesenteric lymphadenitis	3–7 days	ingestion	10^6 cells
Vibrio spp.	intestine	cholera, enteritis	4–96 hours	ingestion	$\sim10^6$ cells
Anaerobes, Gram positive					
Actinomyces spp.	jaw, thorax, abdomen	chronic abcesses, draining sinuses	variable; days or months	direct contact (mouth), airborne, fomites	unknown
Clostridium botulinum	nerves, muscle, skin, intestine (infants)	botulism, acute bilateral cranial nerve impairment	12–36 hours	direct contact, ingestion (non–communicable)	unknown
Clostridium difficile	intestine	pseudomembranous colitis	6–24 hours	environmental	unknown
Clostridium perfringens	skin lesion, large intestine	gas gangrene, food poisoning	1–4 days	ingestion	10^5–10^8 cells
Clostridium tetani	nerves, muscle	tetanus	3–21 days	direct contact (non–communicable)	potent toxin

Table 1. (*Continued*)

Organism	Site(s)	Disease(s)	Incubation period	Mode of transmission	Infectious Dose
Anaerobes, Gram negative					
Bacteriodes spp.	mouth, respiratory track, large intestine	peritonitis, endometritis, abscesses, septicemia	unknown	endogenous	unknown
Bacteria, acid–fast					
Mycobacterium tuberculosis	lower respiratory tract, laryngeal, meningeal	Tuberculosis	4–12 weeks	direct contact, droplet spread, airborne	10 cells
Mycobacterium avium complex	lower respiratory tract, lymph nodes	pulmonary, lymphadenitis	unknown	ingestion, skin lesions (non–communicable)	10^4–10^7 cells
Yeasts					
Candida albicans	mucous membranes, skin	oral thrush, intertrigo, vulvovaginitis, paronychia	variable; 2–5 days in infants	direct contact (sexual), neonatal contact (birth)	unknown
Cryptococcus neoformans	meninges, lower respiratory tract	meningitis, pneumonia	unknown	airborne	unknown
Molds					
Aspergillus spp.	lower respiratory tract	aspergillosis	variable; days to weeks	airborne (non–communicable)	unknown
Blastomyces dermatitidis	lower respiratory tract, skin	blastomycosis	few weeks to months	airborne (non–communicable)	unknown
Coccidioides immitis	lower respiratory tract, skin	coccidioidomycosis	1–4 weeks	airborne (non–communicable)	unknown
Histoplasma capsulatum	lower respiratory tract, skin	histoplasmosis	3–17 days	airborne (non–communicable)	10 spores
Viruses					
Acquired immunodeficiency syndrome	progressive damage to immune and other organ systems	HIV/AIDS	6 months −7+ years	direct contact (sexual, contact with blood or bodily fluids)	unknown
Hepatitis A	liver	hepatitis, type A	10–50 days	ingestion	10–10^3 virus particles
Hepatitis B	liver	hepatitis, type B	24–180 days	direct and indirect contact (bodily fluids), fomites	unknown
Hepatitis C	liver	hepatitis, type C	6–10 weeks	direct contact (contaminated blood)	10^2–10^3 particles/mL blood
Herpes simplex I	skin	herpes (vesicular lesions)	7–10 days	direct contact (saliva)	unknown
Herpes simplex II	skin	herpes (genital)	2–12 days	direct contact (usually sexual)	unknown
Influenza	upper respiratory tract	flu	1–4 days	direct contact, droplet spread, airborne	2–800 virus particles
Measles	skin	rubeola	8–13 days	direct contact, droplet spread	~10 virus particles
Rubella	skin	German measles	12–23 days	direct contact, droplet spread	10–60 virus particles
Varicella–Zoster	skin	chicken pox, shingles	13–17 days	direct contact, droplet spread, airborne	unknown

[†] Infectious Disease Information. (2005, April 29). Infectious disease information, NCID, CDC. [Online]. Centers for Disease Control and Prevention. http://www.cdc.gov/ncidod/diseases/index.htm [2005, August 21]; The "Bad Bug Book." (2003, January 30). FDA/CFSAN Bad Bug Book: Introduction to Foodborne Pathogenic Microorganisms and Natural Toxins. [Online]. U.S. Food and Drug Administration Center for Food Safety and Administration. www.cfsan.fda.gov/~mow/intro.html [2005, August 21].

Infectious Agents MSDS Index. (2003, July 31). Index to Material Safety Data Sheets (MSDS) for Infectious Substances. [Online]. Public Health Agency of Canada. www.phac-aspc.gc.ca/msds-ftss/index.html [2005, August 31].

colony originates from a single cell and is therefore pure of any other cell types. A colony can subsequently be used as a pure sample for microscopy, cell culture, and other analytical tests. Additionally, as only viable cells divide, plating can differentiate between dead microbes and those that are viable and may be the source of infection.

Staining. Microscopic observation of microorganisms is generally preceded by staining a specimen on a microscope slide. The microbe response to various stains (gram-positive/negative, acid-fast, etc.), size, grouping (single, double, chains), and morphology (bacillus, coccus, spirillum, pleomorphic) provide characteristic information helpful in identifying the microorganism. However, several pathogenic species appear similar, or are indistinguishable, under the microscope. For effective observation under a microscope, at least 10^5 cells per milliliter of sample should be present. A colony specimen usually meets this requirement, and preconcentration is generally necessary for viewing a nonplated specimen. Still, very small cells can be difficult to observe and may be overlooked. Finally, microscopic observation usually cannot distinguish between dead and live cells.

Cell Culture. Cell culture is used to ascertain the biochemical properties of a microorganism. A single colony is inoculated into a liquid media broth and incubated. Incubation is usually performed near 37 °C with agitation via shaking or gas sparging to facilitate gas transport into the liquid media for uptake by the microbes. Signs of microbial growth in liquid media include turbidity and gas formation. Turbidity can be used as a simple and nondestructive method to measure cell growth. An optical density measurement provides the degree of light scattering at a particular wavelength through a given path length of liquid media. Increasing cell density due to growth usually increases the degree of light scattered. The measured optical density can, therefore, be directly related to the total cell mass. A calibration curve for each bacterial species is required as various sizes and shapes of different microbes scatter light to varying extents.

Microbial growth in several different media is used to determine a specific microbe's biochemical and physiological characteristics. Definitive identification can require 20 or more media tests. Such tests often use selective media. A media can be made selective by addition of chemicals that inhibit microbe and native flora growth while allowing growth of a specific organism. For example, Thayer–Martin medium selectively isolates pathogenic *Neisseria gonorrhea* and *Neisseria meningitides* (4). The medium contains vancomycin to inhibit growth of gram-positive bacteria, anisomycin to inhibit fungi growth, colistin to inhibit most gramnegative bacilli growth, and trimethoprim-sulfamethoxazole to inhibit *Proteus* growth. The *Neisseria* species are resistant to these inhibitors at the concentrations present in the medium and grow freely.

Antibiotic Susceptibility Testing. Upon isolation and identification of the infectious microbe, antibiotic susceptibility testing can be performed to identify antimicrobial agents that inhibit growth. Additionally, the minimal inhibitory concentration (MIC) is determined by exposing bacteria in media broth to various concentrations of an antimicrobial agent. The lowest antibiotic concentration that inhibits growth is the MIC. A concentration of the antibiotic in the blood at or above the MIC should successfully treat an infection.

Development

Plating, microscopy, cell culture, and susceptibility testing techniques for identifying and treating infectious microorganisms have proven effective against a plethora of pathogens, hence its continued use today. However, these clinical microbiology methods have changed very little over the past century, often require days to obtain confirmed results, and cannot be used successfully to characterize several significant infectious agents including the hepatitis virus. However, with the recent and significant advances in molecular biotechnology, two additional microbe identification methods have found wide use, immunoassay and polymerase chain reaction.

Developed in 1959, the utility of the immunoassay was not fully realized by the medical diagnostic community until the late 1970s and early 1980s. The immunoassay takes advantage of an immune system reaction, the highly specific and strong binding of antibody to antigen. Antibodies are developed that specifically bind a given microorganism, chemical byproducts or proteins produced by a given microorganism, or antibodies produced by the host in response to infection caused by a given microorganism. The developed antibodies are tagged with a reporter molecule. Reporters can be radioisotopes, chemiluminescent or fluorescent molecules, or enzymes (i.e., alkaline phosphatase, horseradish peroxidase) that can produce a radiographic, colorimetric, or fluorescent signal. In the presence of the antigen, the antibody will bind and will remain bound through washes that remove unbound antibody. Detection of the reporter after washes indicates that the antigen was present, as bound antibody was not removed during washing. Although rapid, highly specific, and sensitive, immunoassays cannot differentiate between viable and dead cells and are limited to tests for which antibodies can be developed. They also can be affected by contaminants in the test specimen and do not provide quantitative information regarding the number of pathogenic agents present.

Serological assays are the most commonly used immunoassay in the medical laboratory and by the Centers for Disease Control and Prevention (CDC). The mechanism of Serodiagnosis entails binding of lab-developed antibodies to antibodies produced in the host in response to a specific infection. This indirect method of detecting infectious agents allows identification of microbes that are currently difficult or impossible to isolate and culture. For example, because HIV-1 virus requires advanced containment facilities and is difficult to isolate and culture, it is serologically diagnosed via detection of antibodies produced by the host against the virus. Additionally, a method to effectively isolate and culture hepatitis virus has not yet been devised. Therefore, diagnosis of hepatitis virus infection is done serologically. A lag phase of several weeks often exists between onset of infection and production of antibodies by the host against the microbial agent and, thus, possibly leads to false negatives. False positives are also a concern as antibodies produced by the host during a previous infection may be present and detected.

Immunoassays were the primary diagnostic method used for microbial detection by the CDC until the development of polymerase chain reaction (PCR) (5). PCR is a technique that specifically amplifies DNA sequences (for more information, see page 8). This technology has transformed molecular biology and genetics and has changed diagnostic approaches to the identification, detection, and characterization of infectious agents. With PCR, extremely small quantities of DNA from a microorganism can be amplified and detected. Detection of amplified DNA can occur through gel electrophoresis or via genetic probes. Based upon the highly specific binding between complimentary nucleobases of DNA and RNA, genetic probes are nucleic acid sequences that bind to DNA or RNA unique to a given microorganism. Genetic probes are marked with radioisotopes, chemiluminescent or fluorescent molecules, or enzymes and will produce a quantitative signal only when the complimentary microorganism DNA or RNA is present in the sample (for more information, see page 9). Ou et al. (6) used PCR to amplify and detect HIV sequences from seropositive individuals. Subsequently, PCR amplification and sequence analysis of HIV amplicons (amplified DNA sequences) became the first use of comparative nucleic acid sequence information in a disease outbreak setting (7). Although PCR is very sensitive and sequence analysis provides specific identification capability, these technologies are expensive, time-consuming, labor-intensive, and require expertise in molecular biology. Consequently, use of PCR and genetic probes for identification of microbes is common in research laboratories and academic institutions, but, to date, is not widely used in hospital or medical diagnostic laboratories.

To address rising national and worldwide public health needs, it is desirable that a sensitive, specific, fast, and simple-to-operate device be employed to detect infectious agents. Microbial detection systems that attempt to meet this need have been commercially available since the late 1970s and have progressed significantly with the molecular biotechnology revolution. Still, microbial detection systems face three major challenges: time, sensitivity, and specificity of analysis. Microbial testing and detection must be rapid to allow adequate time for treating the infection and be highly sensitive as a single pathogenic organism may be infectious. Additionally, as a low number of pathogenic microbes may be present in complex biological samples, such as blood or urine, high specificity remains an essential requirement. To tackle these problems, alternative nucleic acid-based approaches have been integrated into user-friendly microbial detection systems that are commercially available for diagnostic purposes.

Contemporary microbial detection systems or biosensors typically consist of a selective biorecognition molecule connected to a transducer that converts a biochemical interaction into a measurable signal. Recognition molecules include nucleic acids, antibodies, and peptides. Commonly used transducers include electrochemical, optical, and piezoelectric. The following sections will discuss numerous commercially available microbial detection systems used in clinical and field settings, including (1) nucleic acid-based, optical technologies and systems; (2) fiber-optic, waveguide-based fluoroimmunoassay systems; (3) a chip- and nanoparticle-based bio-barcode optical technology; (4) an electronic microchip-based technology; and (5) an electronic nose microbial detector.

NUCLEIC ACID-BASED OPTICAL TECHNOLOGIES

Line Immunoprobe Assay (LIPA)

The line immunoprobe assay (LIPA) is a nucleic acid recombinant immunoblotting assay (RIBA) (i.e., oligonucleotides that differentiate different genetic variants are transferred onto a nitrocellulose membrane in a straight line) (8,9). PCR is performed from the clinical sample using primers that selectively amplify a DNA region containing nucleotide differences. The amplicons are hybridized with the immobilized oligonucleotides on the membrane, and an enzyme-based colorimetric method is used to detect binding and positive reactivity. The nucleotide differences contained within the amplified sample DNA provide a unique signature that differentiates target genotypes or mutant microorganisms. These assays were among the first commercially available assays using nucleic acid hybridization for diagnostic purposes.

COBAS AMPLICOR Analyzer

A second commercially available system using PCR technology is the COBAS AMPLICOR Analyzer (Roche Diagnostics; Rotkreuz, Switzerland). This system automates amplification and detection of target DNA from infectious agents by combining five instruments into one: a thermal cycler, automatic pipettor, incubator, washer, and reader. Amplified biotinylated products are captured on oligonucleotide-coated magnetic microparticles and detected colorimetrically with use of an avidin-horseradish peroxidase (HRP) conjugate. The system can detect a broad range of agents including *Bacillus anthracis, Chlamydia trachomatis, Neiserria gonorrhea, Mycobacterium tuberculosis*, cytomegalovirus, hepatitis B and hepatitis C viruses, and HIV in clinical specimens including serum, urine, and sputum. The manufacturer reports that more than 4000 COBAS AMPLICOR Analyzers are currently used in clinical settings worldwide (10).

Real-Time PCR (RT-PCR)

A number of commercially available systems for the diagnosis of infectious diseases make use of a third nucleic acid-based approach (i.e., RT-PCR). As the name implies, RT-PCR, pioneered by Applied Biosystems (Foster City, CA) in the mid-1990s (11), amplifies and measures agent-specific DNA as the reaction proceeds in real-time. It is used to quantify the amount of agent-specific input DNA or cDNA by correlating the amount of DNA with the time it takes to detect a fluorescent signal. This technology uses fluorescent reporter probes (i.e., molecular beacons) that are detected and quantitated at each cycle of the PCR. Molecular beacons are single-stranded, dual-labeled fluorogenic DNA or RNA probes that form a stem loop structure. The loop hybridizes to the target nucleic acid, whereas the stem is end-labeled with a fluorophore at the 5′-end adjacent to a quencher at the 3′-end. Fluorescence resonance energy transfer (FRET) is the process by which energy from an excited fluorophore (donor) is transferred to the adjacent fluorophore (acceptor) at close proximity, resulting in the quenching of fluorescence. Hybridization of the target sequence to the loop separates fluorophore and quencher, and the fluorescence is measured.

The GeneXpert System (Cepheid; Sunnyvale, CA) fully automates and integrates sample preparation with the RT-PCR detection processes. It uses microfluidics technology integrated into disposable assay cartridges. The cartridges contain all the specific reagents required to detect disease organisms such as *Bacillus anthracis, Chlamydia trachomatis*, or foodborne pathogens. The system provides quantitative results from unprocessed clinical samples in 30 minutes or less and is capable is self-cleaning and decontamination before its next use. The GeneXpert module forms the core of the Biohazard Detection System deployed nationwide by the United States Postal Service for anthrax testing in mail-sorting facilities (12). It is also used in hospital laboratories, physician offices, and public health clinics.

Idaho Technology (Salt Lake City, UT) manufactures an automated, field-ready RT-PCR instrument, the R.A.P.I.D. (ruggedized advanced pathogen identification device) system, which is based on the Light Cycler Instrument from Roche Diagnostics (Alameda, CA). The R.A.P.I.D. is developed for military field hospitals and first responders in harsh field environments. Amplification of DNA in real-time can be performed on environmental and blood samples. Idaho Technology claims a 15 minute set-up time and a 20 minute PCR run for a total of 35 minutes using the R.A.P.I.D. Pathogen Test Kit. This instrument is reported to be very sensitive (i.e., *Pseudomonas aeruginosa* was detected in blood culture samples at 10 cfu (colony-forming units)/ml (13). (A colony-forming unit is a single viable cell that forms a colony of identical cells when plated.) This technology is well-established and has been in use worldwide since 1998.

The R.A.P.I.D. technology was recently put to the test at the Prince Sultan Air Base in Saudi Arabia (14). Medical personnel observed a clustering of diarrhea cases and thought them to be due to influenza. However, testing of patient samples with the R.A.P.I.D. identified the cause to be foodborne *Salmonella* within hours of sample submission. Due to the prompt response by medical and services personnel, the outbreak was limited to less than 3% of the base population.

Nucleic Acid Sequence-Based Amplification (NASBA)

A fourth approach for nucleic acid-based detection of infectious organisms is nucleic acid sequence-based amplification (NASBA), a bioMérieux, Inc. (Marcy-l'Etoile, France) proprietary isothermal amplification technology. This method is based on specific amplification of RNA by the simultaneous activity of three RNA-specific enzymes, AMV-reverse transcriptase, T7 RNA polymerase, and RNase H, generating single-stranded RNA as the endproduct (15). NASBA is an isothermal amplification procedure, carried out at 41°C.

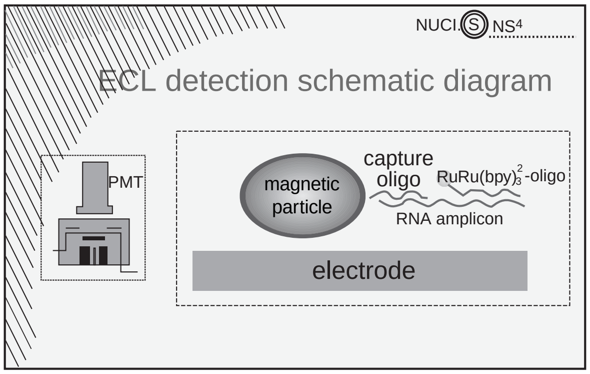


Figure 2. NucliSens Reader ECL detection scheme (16).

Two commercially available systems, NucliSens Reader and NucliSens EasyQ Analyzer (bioMérieux, Inc.), make use of the NASBA technology. Although both systems use NASBA for selective amplification of RNA, as well as a bioMérieux proprietary Boom silica-based nucleic acid extraction method, the NucliSens Reader relies on an electrochemiluminescence (ECL) detection technology, and the NucliSens EasyQ uses fluorescent detection by incorporating specific molecular beacons to which amplicons hybridize. With this method, amplification and detection occur simultaneously in a single tube.

The ECL-based NucliSens Reader employs a sandwich hybridization method for the detection of amplified target RNA (Fig. 2) (16). Two target-specific DNA probes are used: a capture probe bound to magnetic beads and a detection probe labeled with tris (2,2′-bipyridine) ruthenium (Ru). Each of these probes bind to a different region of the target RNA. After the hybridized sample is drawn into the ECL flow cell and the beads are magnetically immobilized on the electrode, a voltage is applied, and the resulting emitted light is detected by a photomultiplier tube (PMT). According to the manufacturer, measurement of 50 reactions takes approximately 50 minutes.

Real-time NASBA and fluorescent detection of target-bound molecular beacons are accomplished by the NucliSens Basic Kit and EasyQ Analyzer (Fig. 3) (17). This technique is most often used to detect RNA viruses. Using a multiplexed NASBA technique to detect four human immunodeficiency virus type 1 (HIV-1) subtypes, DeBaar et al. (18) reported an 89% correct subtype identification relative to sequence analysis and a sensitivity of 92%. The limit of detection was approximately 10^3 copies of HIV-1 RNA per reaction. Lanciotti and Kerst (19) conducted a study comparing TaqMan RT-PCR (Applied Biosystems) and standard reverse-transcription PCR (Roche Molecular Biochemicals) assays with NucliSens NASBA assays

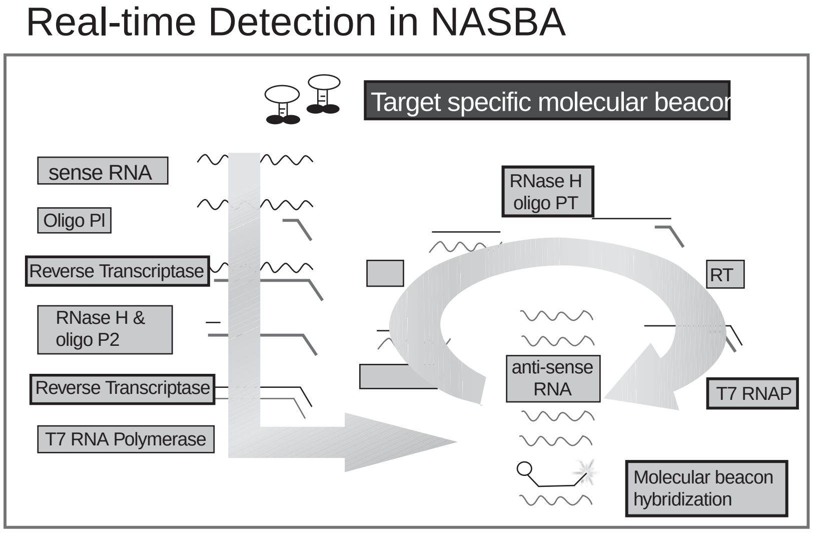


Figure 3. NucliSens EasyQ detection scheme (17).

for detecting West Nile (WN) virus and St. Louis encephalitis virus. The ECL-based and molecular beacon-based NASBA assays demonstrated equal or greater sensitivities and specificities than reverse-transcription PCR in human cerebral spinal fluid. The NASBA-ECL assay for WN virus was 10-fold more sensitive than either the TaqMan or NASBA-molecular beacon assay, detecting 0.01 pfu of WN virus. Moreover, the NASBA molecular beacon-based assay performed significantly faster than either PCR procedures (i.e., a positive signal was detected within 14–45 minutes).

Strand Displacement Amplification (SDA)

A fifth approach for nucleic acid-based detection of infectious organisms is strand displacement amplification (SDA). SDA, first reported by Walker et al. in 1992 (20), is an isothermal process that amplifies DNA or RNA using a restriction enzyme and a DNA polymerase plus several primers, without requiring temperature cycling. Available since 1999, the BDProbeTecET System (Becton, Dickinson & Co.; Franklin Lakes, NJ) couples the proprietary technology, SDA, and real-time fluorescent detection in a rapid one-hour format. This high throughput, chip-based, closed system was developed for detection of *Chlamydia trachomatis* and *Neisseria gonorrhea* in urine samples, endocervical swabs, and male urethral swabs in a one-hour assay time. The complete system includes a sample preparation module, a priming and warming heater unit, and an amplification and fluorescence detection unit. The optical system consists of a fiber-optic bundle with eight branches, and the fluorescent detection reader monitors real-time fluorescence by FRET. Emitted light passes through a custom optical band-pass filter, is detected by a PMT, and is analyzed by software.

Little et al. (1999) (21) reported a sensitivity of 10–15 *N. gonorrhea* cells or *C. trachomatis* elementary bodies. Akduman et al. (22) reported that out of 3544 urine samples tested, 152 were positive using the BDProbeTecET System, and 130 were positive by standard culture techniques resulting in a sensitivity of 99.2% and a specificity of 99.3%.

FIBER-OPTIC FLUOROIMMUNOASSAY SYSTEMS

Analyte 2000 and RAPTOR

The Analyte 2000 and its sister field model, RAPTOR (Research International; Monroe, WA), detection systems use a fiber-optic, waveguide-based sandwich fluoroimmunoassay for the near real-time detection of pathogens in a variety of raw fluid samples (23). Optical fibers are long, thin strands of either glass or plastic that can transmit light over long distances. In the RAPTOR, a monolayer of capture antibodies are immobilized on the surface of a cylindrical waveguide (Fig. 4)(23). The waveguide is incubated with a clinical sample for three to five minutes, washed, and re-incubated with a fluorophore-labeled antibody to form an antibody/antigen/labeled-antibody "sandwich." Excitation light, injected into the waveguide, creates an evanescent wave electric field in the fluid and generates an optical emission from the antibody-antigen complexes. The fluorescent signals are then monitored by a photodetector.

Using the Analyte 2000, the detection limits for *Bacillus anthracis* (vegetative cells) was reported as 30 cfu/ml in water, and for the avirulent strain of *B. anthracis* (i.e., Sterne strain), 100 cfu/ml in whole blood. For spores, the detection limit was 5×10^4/ml (23). The infectious dose of *B. anthracis* in a healthy individual requires inhalation of about ~8,000–50,000 spores (24). This number is reduced in more vulnerable individuals, such as the elderly or those with respiratory problems. Vaccinia virus (a surrogate of the Smallpox virus) from throat swabs was detected at 2.1×10^4 pfu (plaque-forming units, the viral equivalent of bacterial colonies)/ml (25). The infectious dose of smallpox is thought to be low (i.e., 10–100 organisms) (26).

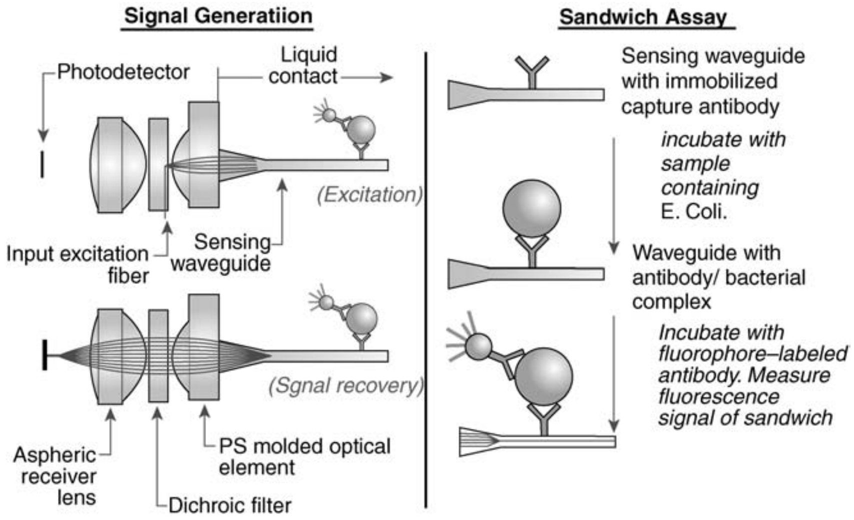

Signal Generatiion
Photodetector
Liquid contact
(Excitation)
Input excitation fiber
Sensing waveguide
(Sgnal recovery)
Aspheric receiver lens
Dichroic filter
PS molded optical element
Sandwich Assay
Sensing waveguide with immobilized capture antibody
incubate with sample containing E. Coli.
Waveguide with antibody/ bacterial complex
Incubate with fluorophore–labeled antibody. Measure fluorescence signal of sandwich

Figure 4. Optical and biomolecular processes of RAPTOR technology (23).

NANOPARTICLE-BASED BIO-BARCODE TECHNOLOGY

Verigene System

The Verigene System (Nanosphere; Northbrook, IL) is an automated device for the chip-based detection of proteins and nucleic acids using an innovative gold nanoparticle-based bio-barcode technology. For proteins, the assay uses two types of probes (Fig. 5 (27): (1) magnetic microparticles (MMPs) functionalized with monoclonal antibodies (mAbs) specific for a target antigen and (2) gold nanoparticles (NP) functionalized with polyclonal antibodies specific for the same target and DNA oligonucleotides (the "bio-barcodes") with a sequence that is a unique identification tag for the target. The Au nanoparticles and the MMPs sandwich the target, generating a complex with a large ratio of barcode DNA to protein target. A magnetic field is applied, allowing the separation of all the MMP/target/NP complexes from the reaction mixture. After a wash to dehybridize the barcode DNA from the nanoparticles, another magnetic field removes the NPs, leaving only the barcode DNA. Detection and identification of the barcodes occurs next through a PCR-less process of amplification. Chip-immobilized capture DNA, complementary with half of the target barcode DNA sequence, is used to bind the barcode DNA. Then, gold nanoparticles, functionalized with oligonucleotides that are complementary to the other half of the barcode DNA, are hybridized to the captured barcode strands. The signal is amplified by the catalytic electrodeposition of Ag onto the Au nanoparticles, and the results are recorded with the Verigene ID system, which measures scattered light intensity from each barcode/Au/Ag complex.

Like protein detection, DNA detection via the nanoparticle bio-barcode approach uses two types of probes (Fig. 6)(28): (1) magnetic microparticles functionalized with oligonucleotides that are complementary to one-half

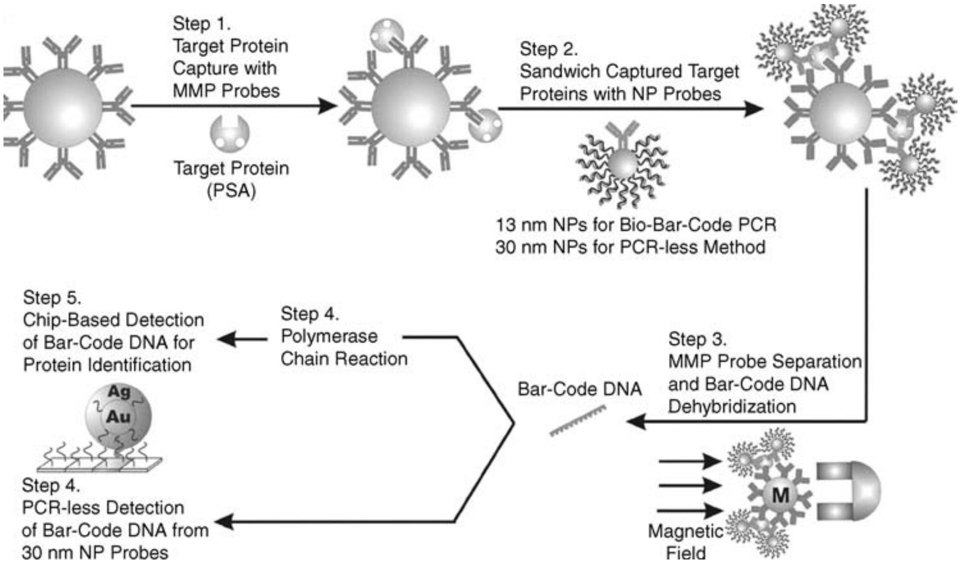

Step 1.
Target Protein Capture with MMP Probes
Target Protein (PSA)
Step 2.
Sandwich Captured Target Proteins with NP Probes
13 nm NPs for Bio-Bar-Code PCR
30 nm NPs for PCR-less Method
Step 3.
MMP Probe Separation and Bar-Code DNA Dehybridization
Bar-Code DNA
M
Magnetic Field
Step 4.
Polymerase Chain Reaction
Step 5.
Chip-Based Detection of Bar-Code DNA for Protein Identification
Ag
Au
Step 4.
PCR-less Detection of Bar-Code DNA from 30 nm NP Probes

Figure 5. Prostate-specific antigen (PSA) detection and barcode DNA amplification and identification (27).

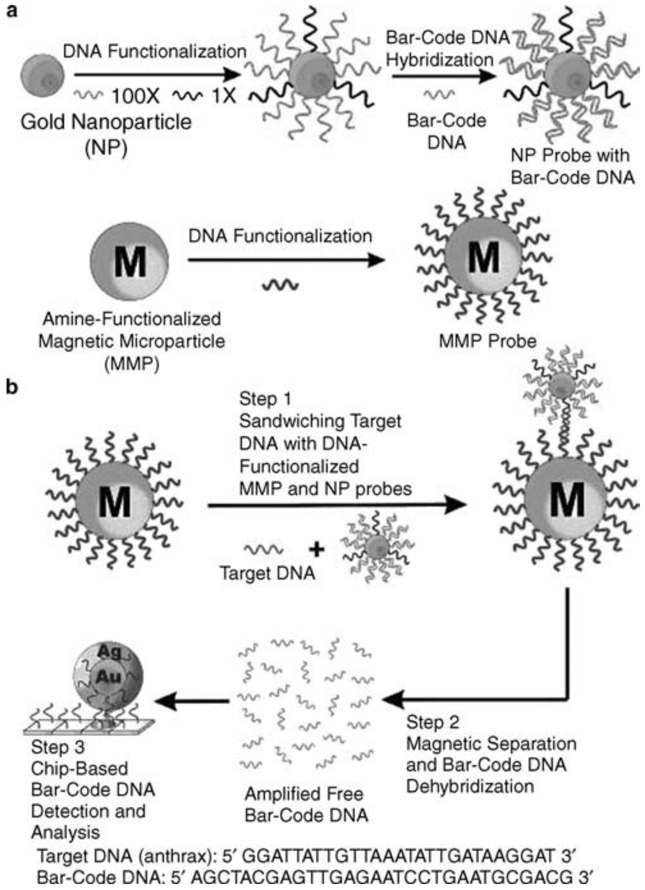


Figure 6. The DNA-bio-barcode assay. (a) Nanoparticle and magnetic microparticle probe preparation. (b) Nanoparticle-based PCR-less DNA amplification scheme (28).

of a target sequence and (2) gold nanoparticles functionalized with two types of oligonucleotides, one that is complementary to the other half of the target sequence and one that is complementary to a barcode sequence that is a unique identification tag for the target sequence. The assay proceeds in the same manner as with protein targets, with the analysis also accomplished by the scanometric method with a Verigene ID system.

The nanoparticle-based bio-barcode approach is reported to provide a sensitivity of 500 zeptomolar, approximately 10 target DNA strands in a 30 μl sample (27). Prostate-specific antigen was detected at 30 attomolar levels with this method, and PCR on the DNA barcodes boosted sensitivity to 3 attomolar (28). The entire assay can be carried out in 3–4 h.

MICROCHIP TECHNOLOGY

NanoChip System

The NanoChip System (Nanogen; San Diego, CA) is an electronic microarray device based on the electrophoretic transport of proteins and nucleic acids on a microchip to specific sites where traditional immunoassays or nucleic acid hybridization reactions occur (Fig. 7) (29). The electronic microchip is a planar array of microelectrodes that electrophoretically transport-charged biomolecules to any individually-electrically-addressed microsite on the surface of the device. Each microsite has an agarose-streptavidin permeation layer coated on top of a platinum microelectrode to bind biotinylated capture molecules. The microchips are referred to as "active electronic microchips" because electric fields are generated for the purpose of transporting biomolecules to and from specific

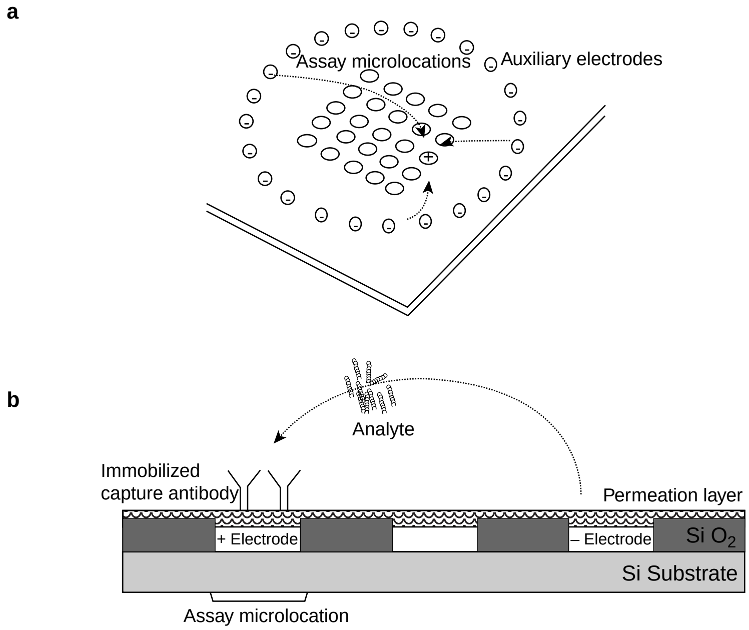


Figure 7. Active electronic microchip technology. (a) Basic chip layout; (b) Cross-section of microchip for electrophoresis of proteins (29).

microsites in a process the manufacturer terms "electronic addressing" (29). As the electric field is generated only in the immediate vicinity of the electrodes, not affecting the solution in other parts of the device, each microsite is an independent assay site allowing for the detection of multiple analytes. Also, the generated electric fields can be used to selectively dehybridize nonspecifically bound analytes from assay sites, which greatly improves the selectivity of the assay. When the biotinylated capture probes are attached to the array, fluorescently labeled analytes are introduced, and further electrical adjustments are made to direct the analytes to concentrate at the microsites for rapid hybridization or antibody-antigen interactions. The fluorescent signal is monitored in a laserinduced fluorescence scanner, and the analytes are identified based on the microlocation of the fluorescence.

Nanogen researchers performed a diagnostic immunoassay for two fluorescently labeled toxins simultaneously, staphylococcal enterotoxin B (SEB) and cholera toxin B. They reported a sensitivity of better than 20 nM concentrations of toxins (29). High specificity was also demonstrated by low nonspecific binding and cross-binding. This assay took 6 minutes to perform, 1 minute for electronic addressing to bind analytes and 5 minutes for washing to reduce nonspecific binding.

More recently, Nanogen researchers reported on an integrated "stacked" microlaboratory for performing automated electric field-driven immunoassays and DNA hybridization assays (30). This device is composed of a CMOS-based electronic microarray chip, a dielectrophoresis microchip, and several modules for DNA sample preparation, strand displacement amplification, and hybridization. *E. coli* bacteria and Alexa-labeled staphylococcal enterotoxin B were detected in the device with specific-to-nonspecific signal ratios of 4.2:1 and 3.0:1, respectively. Identification of the Shiga-like toxin gene from *E. coli* was accomplished in a 2.5 h comprehensive protocol including the dielectrophoretic concentration of intact bacteria, DNA amplification, electronic DNA hybridization to fluorescently-labeled probes, and detection with a fluorescent microscope. This experiment used bacteria cell suspensions of 10^9 cells/ml with a specific-to-nonspecific signal ratio of 22.5:1, showing outstanding specificity.

ELECTRONIC NOSE

Osmetech Microbial Analyzer

An electronic nose is a device that consists of an array of gas sensors with different selectivity patterns, a signal collecting unit, and data analysis by pattern recognition software. When microorganisms grow and metabolize, they emit volatile organic compounds and gases that can be monitored by a biosensor array. The Osmetech Microbial Analyzer (OMA; Osmetech; London, UK) is an automated headspace analyzer using arrays of organic conducting polymers as sensors. The device samples the headspace above the surface of the specimen and detects volatile compounds with an array of up to 48 conducting polymer sensors. Each polymer has unique adsorptive properties, and, once adsorbed, the volatile components modulate the conduction mechanism of the polymer resulting in reversible changes in resistance (Fig. 8) (31). The signal is measured as a percentage change of the original resistance of the polymer. Multivariate data algorithms are used to compare the responses and establish a diagnosis.

When 534 clinical urine samples were analyzed by the OMA, 22.5% had significant bacteriuria (i.e., $>10^5$ cfu/ml), resulting in a sensitivity of 84% and a specificity of 88% relative to standard culture methods (32). Alhough less than optimal, this device shows promise for automated, rapid screening. The company's second FDA approval for detection of bacterial vaginosis was secured in January 2003. Clinical trials with more refined versions of the instrument are in progress. Although electronic nose technology is still in its infancy, it clearly has the potential for providing rapid, sensitive, and simultaneous detection of different strains of bacteria.

FUTURE TRENDS IN MICROBIAL DETECTION SYSTEMS

In the development of the microbial detection systems mentioned, researchers have begun to focus on building integrated devices that combine a pre or post-processing step such as PCR-based amplification, with post-derivatization (fluorescent labeling) and detection. Microarray technologies are being developed to overcome limitations of sample volume and high throughput analysis. In the

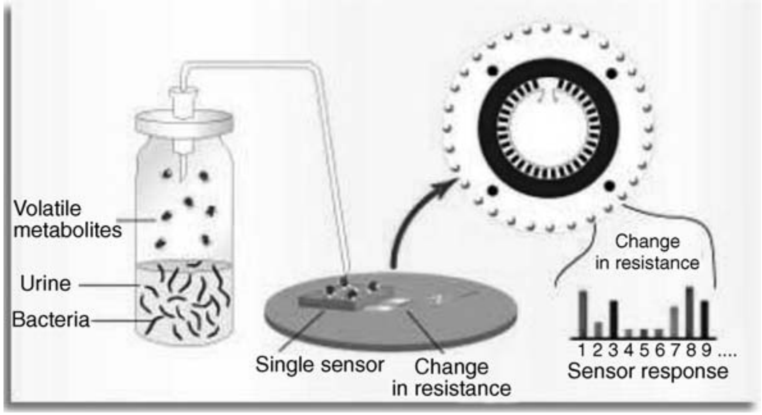


Figure 8. Osmetch Microbial Analyzer detection technology (31).

near future, biomedical science can realize the integration of all laboratory equipment used in molecular biology on a chip-based platform arrayed to detect large numbers of pathogens in a high throughput, portable device. Biological Micro-Electro-Mechanical systems (BioMEMS), also referred to as lab-on-a-chip and micro total analytical systems (μ-TAS), is an area rapidly advancing due to the integration of micro and nanotechnology with biotechnology. Current reviews for this technology are abundant in the literature (33,34). Microfluidic-based devices have been on the market since 1999, but much work is still underway to build modular-type systems with complete integration of sample collection, concentration, pre and post-processing steps, separation, selective capture, viability detection, lysing, and protein and DNA analysis. Development of such systems in a high throughput fashion capable of detecting and discriminating between hundreds of pathogenic agents would impact not only medical diagnostics but homeland security and public health, including home monitoring, medicine, and veterinary diagnostics. As the field moves toward lab-on-a-chip systems, cost, limited sample throughput, ease-of-use, and limited waste production (reagentless systems) will be considered in design strategies. The second progression toward advanced microbial detection systems will be the incorporation of nanotechnology. Current nanotechnologies such as quantum dots, nanoparticles, and synthetic nanopores are already being incorporated into current chip-based diagnostic systems. CellTracks technology (Immunicon Corporation, Huntingdon Valley, PA) has developed magnetic nanoparticles called ferrofluids, which consist of a magnetic core encompassed by a polymer coating tagged with antibodies for whole cell and pathogen detection. Up-Converting Phosphor Technology (UPT), by OraSure Technologies, Inc., makes use of proprietary ceramic nanoparticles for DNA detection. These particles have been shown to be a 1000 times more sensitive than fluorescent technologies. Finally, a trend exists to build detection systems from the bottom up rather than the top down. Small building blocks such as protein motors are being designed to move cargo including peptides and antibody fragments as a method of patterning arrays. "Switchable" materials such as poly-n-isopropylacrylamide (PNIPAM) are used to pattern antibodies, capture proteins, and move fluids, replacing mechanical components of BioMEMS systems. PNIPAM has a thermally activated lower critical solubility temperature (LCST) of 32 °C. At temperatures below the LCST, the polymer swells in water to create a hydrophilic surface that resists protein adsorption. Above the LCST, the polymer collapses to form a hydrophobic surface that promotes protein adsorption. Whether the bottom up approach based solely on nanomaterials will hold in the long run remains to be seen. However, it is clear that nanotechnology that complements and extends current MEMS detection methods will revolutionize the field of medical diagnostics. Early examples of lab-on-a-chip technologies integrating nanotechnologies already exists, which address the current limitations of detection systems. The approach is toward development of portable microsystems that are reagentless; handle small sample size; eliminate the need of labels and probes; are specific, sensitive, and high throughput; perform multiple functions from sample concentration to final detection; and are easy to use.

Briefly, some of these technologies include cantilever arrays, which operate by a slight bending of the cantilever beam at the nanoscale level upon analyte binding. Protiveris, Inc. (Rockville, MD) is developing microcantilever arrays for combined detection of DNA and protein. Capture molecules are attached to the beams and, as samples moves across the device, binding of a target molecule results in nm bending of the beam. These devices can be integrated into microfluidic systems, require no labels or reagents, and are very sensitive and specific. Nanowires, nanoneedles, and nanoelectrode arrays are additional technologies that can detect multiple analytes simultaneously. Electronic signals can be averaged over thousands of electrodes eliminating the need for PCR amplification, and no reagents are required. These devices are coated with selective molecular recognition molecules and change in conductance occurs during a binding/recognition event. Aside from integration into lab-on-a-chip systems, these technologies have applications in *in vivo* medical diagnostics.

Over the next few years, nanotechnologies will continue to evolve and become integrated into chip-based microsystems for detection, diagnostics, and drug delivery. Later, in perhaps 20– 30 years, the introduction of nanomachines for *in vivo* diagnostics and treatment may well emerge, changing the current way of conducting medical and healthcare practice.

BIBLIOGRAPHY

1. Gannon JC. The Global Infectious Disease Threat and Its Implications for the United States. [Online]. Federation of American Scientists. http://www.fas.org/irp/threat/nie99-17d.htm.
2. Fauci AS. Global Health: The United States Response to Infectious Diseases, Testimony before the U.S. Senate Labor and Human Resources Subcommittee on Public Health and Safety. [Online]. National Institute of Allergy and Infectious Diseases, National Institutes of Health. http://www3.niaid.nih.gov/about/directors/congress/1998/0303/default.htm.
3. Ivnitski D, Abdel-Hamid I, Atanasov P, Wilkins E. Biosensors for detection of pathogenic bacteria. Biosens Bioelectron 1999;14:599–624.
4. Washington JA. Principles of Diagnosis. [Online]. University of Texas Medical Brach. http://gsbs.utmb.edu/microbook/ch010.htm.
5. Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB, Erlich HA. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 1988;239:487–491.
6. Ou CY, Kwok S, Mitchell SW, Mack DH, Sninsky JJ, Krebs JW, Feorino P, Warfield D, Schochetman G. DNA amplification for direct detection of HIV-1 in DNA of peripheral-blood mononuclear-cells. Science 1988;239:295–297.
7. Ou CY, Ciesielski CA, Myers G, Bandea CE, Luo CC, Korber BTM, Mullins JI, Schochetman G, Berkelman RL, Economou AN, Witte JJ, Furman LJ, Satten GA, Macinnes KA, Curran JW, Jaffee HW. Molecular epidemiology of HIV transmission in a dental practice. Science 1992;256:1165–1171.
8. Stuyver L, Van Geyt C, De Gendt S, Van Reybroeck G, Zoulim F, Leroux-Roels G, Rossau R. Line probe assay for monitoring

drug resistance in hepatitis B virus-infected patients during antiviral therapy. J Clin Microbiol 2000;38:702–707.

9. Stuyver L, Wyseur A, Vanarnhem W, Lunel F, Laurentpuig P, Pawlotsky JMM, Kleter B, Bassit L, Nkengasong J, Vandoorn LJ, Maertens G. Hepatitis C virus genotyping by means of 5′-UR/core line probe assays and molecular analysis of untypeable samples. Virus Res 1995;38:137–157.
10. Roche Diagnostics - Roche Molecular Diagnostics - Products. Automated PCR Systems. [Online]. Roche Molecular Diagnostics. http://www.rochediagnostics. com/ba_rmd/rmd_ products_automated_pcr_systems23.html.
11. Heid CA, Stevens J, Livak KJ, Williams PM. Real time quantitative PCR. Genome Res 1996;6:986–994.
12. Cepheid – Products. (No date). GeneXpert Technology, System Overview. [Online]. Cepheid. http://www.cepheid.com/Sites/cepheid/content.cfm?id=164.
13. Jaffe RI, Lane JD, Bates CW. Real-time identification of *Psuedomonas aeruginosa* direct from clinical samples using a rapid extraction method and polymerase chain reaction (PCR). J Clin Lab Analysis 2001;15:131–137.
14. Idaho Technology - R.A.P.I.D. Successes using the R.A.P.I.D. [Online]. Idaho Technology. http://www.idahotechnology.-com/rapid/success.htm.
15. NucliSens Key Technologies. (2005). bioMérieux – Clinical Microbiology Products. [Online]. bioMérieux, Inc. http://www.biomerieuxusa. com/clinical/nucleicacid/technology. htm.
16. NucliSens Reader. (2005). bioMérieux – Clinical Microbiology Products. [Online]. bioMérieux, Inc. http://www.biomerieux-usa.com/clinical/nucleicacid/reader.htm.
17. NucliSens EasyQ. (2005). bioMérieux – Clinical Microbiology Products. [Online]. bioMérieux, Inc. http://www.biomerieux-usa.com/clinical/nucleicacid/easyq/easyq_technology.htm.
18. DeBaar MP, Timmermans EC, Bakker M, Rooij E, van Gemen B, Goudsmit J. One-tube real-time isothermal amplification assay to identify and distinguish human immunodeficiency virus type 1 subtypes A, B, and C and circulating recombinant forms AE and AG. J Clin Microbiol 2001;39:1895–1902.
19. Lanciotti RS, Kerst AJ. Nucleic acid sequence-based amplification assays for rapid detection of West Nile and St. Louis encephalitis viruses. J Clin Microbiol 2001;39:4506–4513.
20. Walker GT, Little MC, Nadeau JG, Shank DD. Isothermal *in vitro* amplification of DNA by a restriction enzyme/DNA polymerase system. Proc Natl Acad Sci USA 1992;89:392–396.
21. Little MC, Andrews J, Moore R, Bustos S, Jones L, Embres C, Durmowicz G, Harris J, Berger D, Yanson K, Rostkowski C, Yursis D, Price J, Fort T, Walters A, Collis M, Llorin O, Wood J, Failing F, O'Keefe C, Scrivens B, Pope B, Hansen T, Marino K, Williams K, Boenisch M. Strand displacement amplification and homogeneous real-time detection incorporated in a second-generation DNA probe system, BDProbeTecET. Clin Chem 1999;45:777–784.
22. Akduman D, Ehret M, Messina K, Ragsdale S, Judson FN. Evaluation of a strand displacement amplification assay (BD ProbeTec-SDA) for detection of *Neisseria gonorrhoeae* in urine specimens. J Clin Microbiol 2002;40:281–283.
23. RAPTOR. RAPTOR, Portable, Multianalyte Bioassay System. [Online]. Research International. http://www.resrch-intl.com/raptor.html.
24. Bacillus anthracis. (June 2, 2003). *Bacillus anthracis* (anthrax). [Online]. http://microbes.historique.net/anthracis.html.
25. Donaldson KA, Kramer MF, Lim DV. A rapid detection method for Vaccinia virus, the surrogate for smallpox virus. Biosens Bioelectron 2004;20:322–327.
26. Franz DR, Jahrling PB, Friedlander AM, McClain DJ, Hoover DL, Bryne WR, Pavlin JA, Christopher GW, Eitzer EM, Jr.. Clinical recognition and management of patients exposed to biological warfare agents. JAMA 1997;278:399–411.
27. Nam J-M, Thaxton CS, Mirkin CA. Nanoparticle-based bio-barcodes for the ultrasensitive detection of proteins. Science 2003;301:1884–1886.
28. Nam J-M, Stoeva SI, Mirkin CA. Bio-bar-code-based DNA detection with PCR-like sensitivity. J Am Chem Soc 2004;126:5932–5933.
29. Ewalt KL, Haigis RW, Rooney R, Ackley D, Krihak M. Detection of biological toxins on an active electronic microchip. Anal Biochem 2001;289:162–172.
30. Yang JM, Bell J, Huang Y, Tirado M, Thomas D, Forster AH, Haigis RW, Swanson PD, Wallace RB, Martinsons B, Krihak M. An integrated, stacked microlaboratory for biological agent detection with DNA and immunoassays. Biosens Bioelectron 2002;17:605–618.
31. Osmetech. (2005). eNose Technology. [Online]. Osmetech. http://www.osmetech.plc.uk/enose.htm.
32. Aathithan S, Plant JC, Chaudry AN, French GL. Diagnosis of bacteriuria by detection of volatile organic compounds in urine using an automated headspace analyzer with multiple conducting polymer sensors. J Clin Microbiol 2001;39:2590–2593.
33. Bashir R. BioMEMS: State-of-the-art in detection, opportunities and prospects. Adv Drug Delivery Rev 2004;56:1565–1586.
34. Lee SJ, Lee SY. Micro total analysis system (μ-TAS) in biotechnology. Appl Microbiol Biotechnol 2004;64:289–299.

See also COLORIMETRY; COMPUTER-ASSISTED DETECTION AND DIAGNOSIS; COMPUTERS IN THE BIOMEDICAL LABORATORY; FLUORESCENCE MEASUREMENTS; SAFETY PROGRAM, HOSPITAL.

MICROBIOREACTORS

XIAOYUE ZHU
TOMMASO BERSANO-BEGEY
YOKO KAMOTANI
SHUICHI TAKAYAMA
University of Michigan
Ann Arbor, Michigan

INTRODUCTION

This article defines microbioreactors as micrometer scale reaction vessels in which biological reactions are performed. Whereas large-scale bioreactors mainly focus on efficiently producing desired end products, microbioreactors have additional targeted applications such as studying cellular processes under simulated physiological microenvironments, and functioning as portable cellular biosensors or implantable devices inside the body to restore tissue functions. Increasing needs in developing personalized cell-based therapies and medicines together with rapid advances in micro- and nanotechnologies ensure that the field of microbioreactors will continue to flourish. Although most microbioreactors are still in their infancy, relatively simple compared to their macroscopic counterparts, and often not fully integrated or packaged into compact self-contained platforms, they already have started to have an

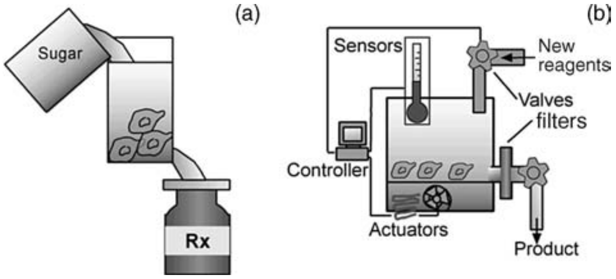


Figure 1. What are bioreactors and microbioreactors? Part (a) shows a generalized bioreactor application, feeding cells to produce complex biomolecules for pharmaceutical applications. Part (b) shows the main components of both bioreactors and microbioreactors: a reaction chamber containing main reagents and cells, sensors (for temperature, flow rate, pH, oxygen, glucose, etc.), actuators (pumps, valves, heaters, mixers, filters), and controllers to regulate actuators based on sensor readings.

impact in pharmaceutical, medical, clinical, and biological fields (Fig. 1).

Microbioreactors enable low cost high throughput investigation of bioreactions in ways not possible with macroscopic bioreactors. For example, the development of a biopharmaceutical requires extensive optimization processes prior to scale-up for industrial manufacture: The bacterial strain or mammalian cell that produces the highest yield of a particular product is constructed or identified; cultural parameters (e.g., temperature, pH, oxygen concentration, and media composition) are systematically evaluated to maximize cell growth and product formation; the design of the reactors themselves are adjusted iteratively as the development process gradually scale-up from benchtop to production scale. With microfabrication techniques, hundreds and thousands of small reaction chambers can be produced simultaneously and operated in parallel allowing vast combinations of parameters to be screened concurrently in short periods of times (Fig. 2a). The consumption of reagents and resources are reduced dramatically because for a given reaction chamber geometry, every 10-fold reduction in linear dimension would lead to a 1000 time reduction in its volume.

To enhance the quality of information obtained from cellular studies, rather than simply increasing throughput, some microbioreactors are designed to simulate physiological cellular microenvironments. Using microbioreactors for studying cellular physiology makes intuitive sense when one considers that much of the bioreactions within living organisms occur at the microscale. For example, capillary blood vessels, lung small airways, livers sinusoids, kidney nephrons, and reproductive tracts are all networks of small sacs, ducts, and tubes. Cells in these and other similar systems are constantly perfused with nutrients and oxygen and exposed to shear stresses and gradients of chemicals (1–3). Conventional *in vitro* cell culture studies, such as culture dishes or 96 well plates, often fail to present many of these dynamic physiological parameters. Microbioreactors, however, can be designed to simulate many such physical and chemical conditions that cells experience inside the body. In the body, different tissues and cell types also interact and communicate with each other. Microbioreactors can be designed to network multiple reaction chambers together to capture the complexity of living organisms and used as animal and human surrogates in pharmacokinetic and toxicology studies. Microbioreactors can also work as cell-based biosensors, where effects on cell behaviors serve as readouts for selective and sensitive detection (Fig. 2c).

Some microbioreactors aim to continuously produce and deliver small quantities of biopharmaceuticals (e.g., insulin for regulation of blood sugar) inside the body (Fig. 2e). Implantable devices that continuously produce drugs based on physiological demands would eliminate the need for repeated injections and blood tests, and allow for delivery of stable, safe, and effective doses that resemble physiological concentration profiles. Such devices contain cells that use nutrients from the body to produce and secrete drugs (4).

Besides cell-based microbioreactors, enzymatic microbioreactors are also useful for detection and analysis. Enzyme-linked immunosorbent assays (ELISA), for example, are useful in high sensitivity detection and analysis of a wide variety of proteins and chemicals related to pollution, disease, and basic biology. Microscale systems often require shorter incubation time due to the short distances that analytes and reagents have to diffuse. Microbioreactors are also more portable and thus suitable for field and point-of-care use.

The development, integration, and packaging of microbioreactors present many challenges, as well as unique opportunities. Because of scaling issues and fabrication limitations, functional microscopic devices often require totally new designs instead of simply “shrinking” their macroscopic counterparts. Thus microscale pumps, valves, and mixers not only look different from their macroscopic counterparts, but may also operate with different mechanisms. Microscale sensors, another crucial component of microbioreactors, is an active area of research that has yielded a variety of useful systems based on optical, electrochemical, ultrasonic, and mechanical detection schemes. Although many microscale components have been developed to date, functional microbioreactors that integrate multiple pumping, valving, mixing, sensing, and control features are still relatively uncommon. Microbioreactors generally have fewer components integrated into their systems compared to their larger counterparts.

The organization of the remainder of this article is as follows. First, the principles that govern microscale reactions and microbioreactor operation are discussed. Second, elements of microbioreactor components are described, along with a brief description of the microfabrication processes that can be used to create them. Finally, highlights of state-of-the-art microbioreactors are given with focuses on production optimization, clinical treatment, toxicology testing, development of implantable systems, and basic cell biology studies. Because of space limitations, this article is representative rather than comprehensive. We also focus mainly on cell-based bioreactors rather than enzyme-based ones. Interested readers are referred to reviews on immobilized microfluidic enzymatic reactors (5). We also exclude important bioreactor categories that are not directly

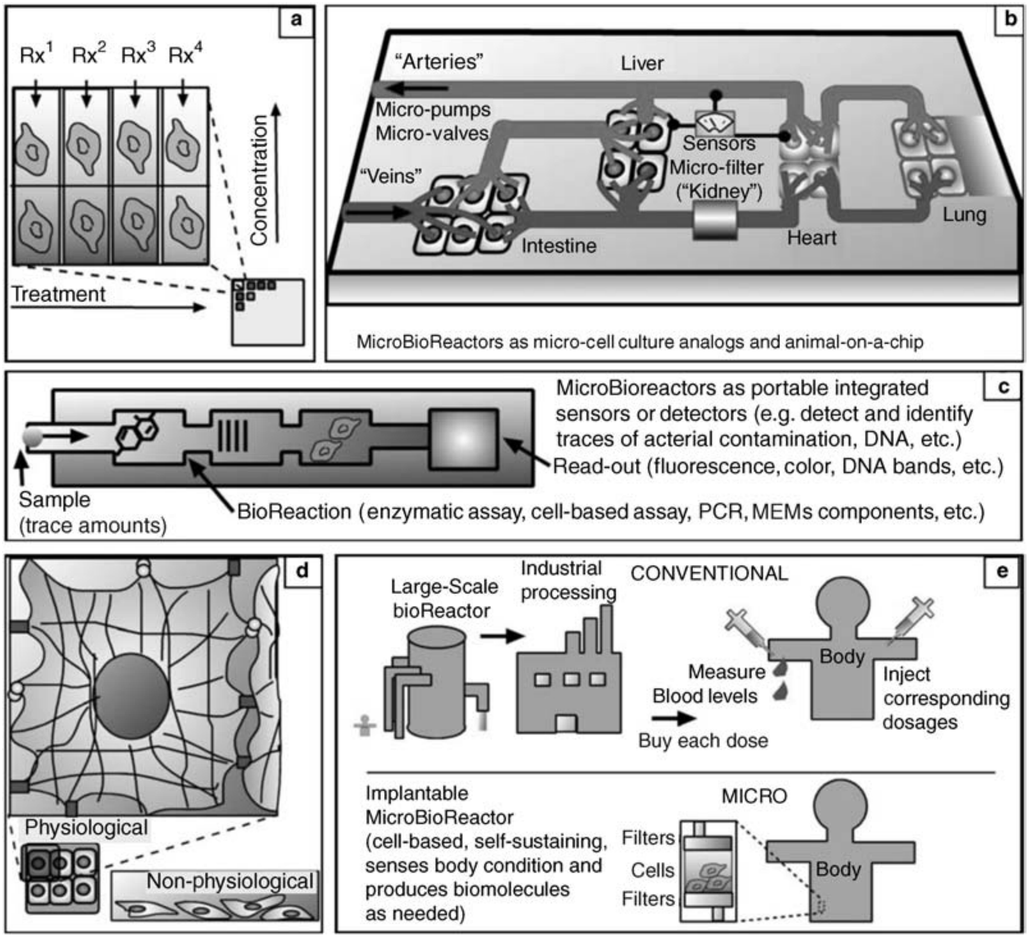

a
Rx1 Rx2 Rx3 Rx4
Concentration
Treatment
b
"Arteries"
Liver
Micro-pumps
Micro-valves
"Veins"
Sensors
Micro-filter
("Kidney")
Intestine
Heart
Lung
MicroBioReactors as micro-cell culture analogs and animal-on-a-chip
c
MicroBioreactors as portable integrated sensors or detectors (e.g. detect and identify traces of acterial contamination, DNA, etc.)
Read-out (fluorescence, color, DNA bands, etc.)
Sample
(trace amounts)
BioReaction (enzymatic assay, cell-based assay, PCR, MEMs components, etc.)
d
Physiological
Non-physiological
e
Large-Scale bioReactor
Industrial processing
CONVENTIONAL
Measure
Blood levels
Body
Inject corresponding dosages
Buy each dose
Implantable MicroBioReactor (cell-based, self-sustaining, senses body condition and produces biomolecules as needed)
MICRO
Filters
Cells
Filters
Body

Figure 2. Advantages and new applications of microbioreactors over large scale bioreactors: (a) microscale parallel cell culture and testing in a microchip configuration can be used to test and compare hundreds of conditions and parameters simultaneously, to optimize a reaction or screen potential treatments. (b) Microdevices containing interconnected chambers and microcultured cells from different tissues can be used to test drug effects on entire organisms, taking into account systemic impact of organs such as liver on drug metabolism, possibly replacing some stages of animal and human testing. (c) Microbioreactors can be used as detectors for bioactive compounds (e.g., toxins, endocrine disruptors, drugs) or biodisruptive conditions (e.g., radiation) with the advantage that all the steps of a test can be integrated in a single device that requires only trace amounts of samples and reagents. Cell-based sensors may be able to detect unknown reagents that affect living organisms. (d) Microfabrication can also provide microbioreactors with more physiologically accurate microenvironments, thus making *in vitro* testing more reliable and closer to *in vivo* conditions: for example, in actual tissues, each cell is held in its three-dimensional (3D) shape by tension through connected cytoplasmic fibers (the cytoskeletons), and cells have many surface interactions with other surface microfeatures and other cells. In contrast, in conventional cell culture that cannot microfabricate these features, cells configuration, and environment conditions, such as the amount of surface contacts, are drastically different. (e) Implantable devices containing cells such as insulin-producing pancreatic islets can be used as an improved treatment that continuously produce insulin based on the body's minute-by-minute needs, eliminating the need for periodical blood tests and consequent injections of insulin mass-produced in conventional bioreactors.

related to medicine, such as those used for food testing, plant cell culture, and wastewater treatment.

MICROBIOREACTOR DESIGN PRINCIPLES

Basic principles of bioreactor operations, such as thermodynamics, kinetics, mass transfer, sterilization, and structural considerations, are similar for both macro- and microbioreactors. The implementation of these principles into actual practice, however, can differ greatly.

Scaling Effects

At first glance, development of microbioreactors may seem to be a matter of simply shrinking macroscopic components

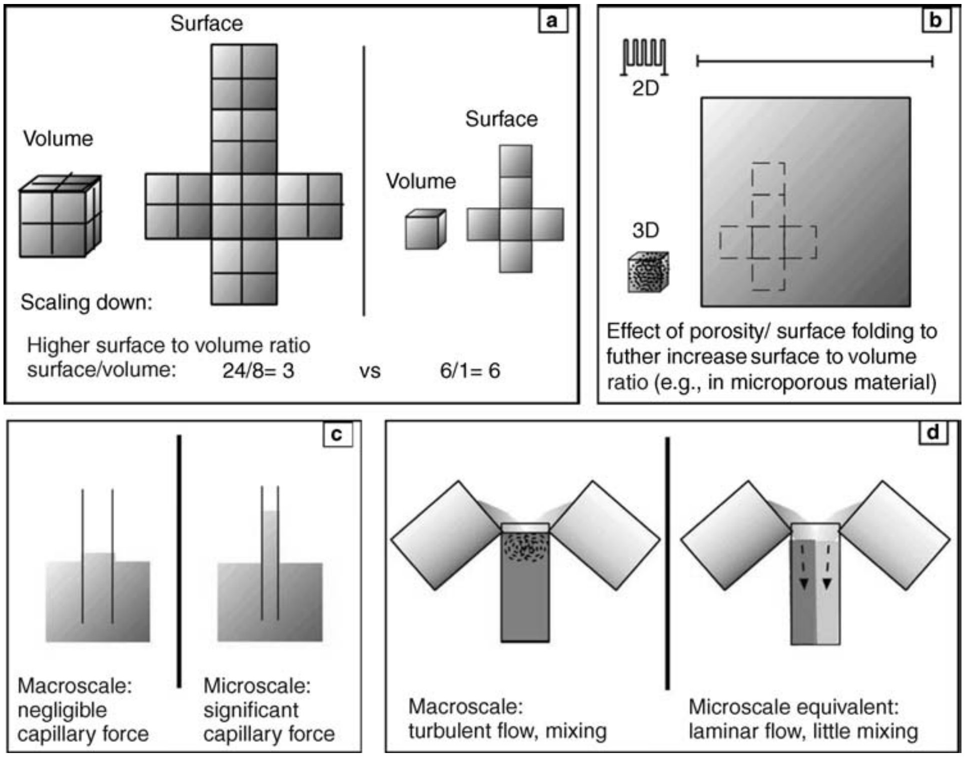

a
Surface
Volume
Scaling down:
Higher surface to volume ratio
surface/volume: 24/8= 3 vs 6/1= 6
Surface
Volume
b
2D
3D
Effect of porosity/ surface folding to futher increase surface to volume ratio (e.g., in microporous material)
c
Macroscale: negligible capillary force
Microscale: significant capillary force
d
Macroscale: turbulent flow, mixing
Microscale equivalent: laminar flow, little mixing

Figure 3. Microscale physics useful for microbioreactors. (a) As dimensions of devices scale down to microscopic scales, surfaces and volumes scale at different rate, so that smaller devices have a much greater surface/volume ratio. This change affects many physical phenomena such as capillary force (c) and produces effects such as laminar flow (d), which allows two liquids to flow side by side with minimal turbulent mixing. These microscale phenomena can be useful for microbioreactor opera tion. For example, capillary force can be used to pump liquid through microchannels, and laminar flows can be used to selectively deliver different reagents to different parts of single cells. (b) Microfabrication can further increase the surface/volume ratio, in a biomimetic manner, by creating microporous structures. This is useful since many biological reactions occur on surfaces.

into smaller ones. Directly downsizing systems and components, however, compromise their functions. Operations at microscales are often altered in unexpected ways compared to those at macroscales because the relative importance of physical effects that determine the functions of the microbioreactor components is size dependent. For example, if the linear dimensions of an object are reduced equally by a factor of 100, the surface area decreases by a factor of 10,000, and the volume decreases by a factor of 1,000,000 (Fig. 3a). The resulting increase in surface/volume ratio leads to prominence of surface effects such as surface tension and viscous drag over gravity and momentum. Thus, at the microscale, laminar flow, diffusive mixing, and capillary forces dominate over turbulence, convective mixing, and gravitational forces (Fig. 3c, d) (6). Scaling laws affect almost every aspect of microbioreactor operation. For example, it is more difficult to perform turbulent mixing at the microscale compared to at the macroscale (7–9), but more convenient to use electroosmosis flow to transport fluid samples through microchannels (10). More details of how scaling laws affect the design and operation of the microbioreactor components are noted in each section below.

Microbioreactor Operation Principles

The main challenge in bioreactor operation lies in its dynamic nature. Optimal reaction conditions must be maintained even as multiple parameters, such as concentration of molecules, activity of enzymes, quantity and quality of cells, and heat production are changing. Understanding the reaction mechanisms is helpful for optimizing reactor designs. It is also desirable to constantly monitor and control reaction environments in real time.

Mass And Heat Balance. In bioreactor processes, it is common to assume steady state, that is, a balance of input, output flow of materials, and thermal homeostasis. As substrate input is converted into product, mass balances need to be closely monitored and controlled. In many enzymatic processes, for example, the reactions are reversible or are inhibited by products. In such cases, it is beneficial to favor the forward enzymatic reaction by keeping the upstream reagent concentrations high and constantly removing downstream products. This type of balance can be achieved to different extents by fed-batch systems or continuous flow systems that constantly replenish and remove reactor contents. Even in so-called batch reactors, where there is no flow of liquid in and out of the reaction chamber, it is often crucial to aerate and maintain sufficient levels of oxygen at all times.

As dimensions are diminished, the conductive resistance at the channel wall/fluid interface decreases, and temperature gradient increases, leading to a greater efficiency of thermal conductivity (11). Because heat flux scales with surface area while heat capacity scales with volume, thermal time constants and heat flux scale linearly with decreasing dimensions. The small thermal mass and associated fast thermal response allows for quick temperature control. On the other hand, since the microsystems can be heated and cooled quickly, it is difficult to maintain a constant temperature; external temperatures must be set at the desired temperature, or constant heating and cooling of the microbioreactor will be required to avoid fluctuations.

A phenomenon that may not impact macroscopic bioreactors considerably, but has a significant effect on microbioreactors, is evaporation. Unless carefully monitored, evaporation will greatly affect reactant concentrations and osmolarity because of the small volume that a microbioreactor can hold.

A variety of mathematical models have been developed to analyze and control bioreactor thermodynamics and stoichiometry. Interested readers are referred to existing texts on the topic (12). For example, Roels (13,14) developed correlations to estimate some of the important thermodynamic parameters to predict yields, substrate and oxygen requirements, and heat dissipation for cellular reactions.

Reaction Kinetics. Bioreaction kinetics depends critically on temperature, pH, dissolved oxygen concentrations, and presence of inhibitors or enhancers. Oxygen fluctuation affects the metabolic and signaling pathways of cells. For bacterial bioreactors, the main concern is often to achieve a high enough oxygen transfer rate to match the oxygen uptake rate and maintain sufficiently high dissolved oxygen concentrations. For mammalian cell cultures, it is necessary to maintain an appropriate dissolved oxygen concentration. In expansion of stem cells, for example, lower oxygen concentrations that mimic physiological values gives higher yields and purity of cells (15,16). Microbioreactors present unique challenges and opportunities in terms of oxygen transfer. Microbial and eukaryotic cell cultures require constant replenishment of dissolved oxygen into the medium because of the low solubility of oxygen in aqueous solutions (8 $\mathrm{mg/L^{-1}}$ at 35 °C in distilled water) (17). For larger bioreactors, bubbling oxygen or passing ambient gas through the cell culture suspension are commonly used (18). Microscale gas formation is more challenging, but has been demonstrated elegantly using electrolysis by Maharbiz et al. (19) (see Components section for details). For microbioreactors with submilliliter volumes, bubbles are generally avoided because they can clog channels and are difficult to dislodge from microchannel walls. For such systems, it is common to use a gas-permeable membrane to oxygenate the culture medium (20,21). The large surface/volume ratio of microdevices provides an advantage for such membrane-based oxygen transfer processes. Poly(dimethylsiloxane) (PDMS), a biocompatible polymer material commonly used for microdevice fabrication, is particularly advantageous in this regards due to its high permeability to oxygen.

Enzyme performance and cell growth and function are particularly sensitive to pH changes. When the reaction is not within the optimal pH range, the reaction rate declines dramatically (12). Substrates, products, or contaminants can reversibly or irreversibly affect enzyme activity. Surface interactions can also alter the kinetics of enzyme reactions, stability of enzymes (22), as well as growth and function of cells (23). Precisely dispensing nano- or picoliters of acids and bases into the tiny reaction chamber to adjust pH is technically challenging. Furthermore, because it is difficult to perform convective mixing in microbioreactors, there may be local variations in the pH and in the reaction kinetics.

Mechanical Considerations. When microbioreactors are coupled with microfluidic systems, cells are subjected to shear stresses. For a given maximum flow velocity, smaller channels will give rise to larger shear stresses (1). Although excess forces can damage cells, moderate shear stresses on genetically engineered cells have actually enhanced the production yield of recombinant proteins (24). Cells in the body are also often exposed to stretching and/or compression. These forces are critical factors that regulate function of cells in muscles, hearts, blood vessels, lungs, and other tissues (25,26). For example, shear stress regulates morphology, gene expression and function of endothelial cells, the monolayers of cells lining the inside of blood vessels (27–30). Shear stress is physiologically important for maintaining vascular homeostasis (31), regenerate bone and healing fractures (32), differentiate embryonic stem cells into cardiovascular fate, leukocytes rolling and tethering to endothelial cells (33). Nuclear factor-κβ (NF-κβ), for example, has been identified as a shear stress- responsive transcription factor that enhances the transcription of many genes including cytokines, growth factors, adhesion molecules, and immunoreceptors in response to shear stress (34).

Some of the biochemical and structural responses of cells to stretch include enhanced expression of endothelin (35), nitric oxide (36), and integrin (37) in vascular endothelial cells, and increased extracellular matrix production in cardiac fibroblasts (38) and in smooth muscle cells (39). Mechanical loading influence cellular functions *in vitro*, including proliferation (40,41), differentiation (42), hypertrophy (40,43), alignment (41,44), G protein activation, second messenger activity (45), and gene expression (43,46).

Cells in the body are constantly subjected to physical forces, such as cyclic mechanical deformation involving tension, compression, shear stress, or all three. Cell proliferation, differentiation, migration, signal transduction, and gene expression are all affected by such mechanical forces (47). Although it is largely unknown how mechanical stimuli are converted into intracellular signals of gene expression (48), there are many efforts to recreate physiological mechanical signals *in vitro*, either in the form of shear stresses from fluid flow, hydrodynamic pressures, or mechanical stresses applied to cells through the substrates. An ideal microbioreactor would provide optimal biomechanical and biodynamic stimuli for cell and tissue growth.

Material Considerations. The material used to fabricate the reaction chambers is crucial because they directly contact cells, enzymes, and products of bioreactors. Sometimes, it is also necessary to mimic physiological cellular environments (Fig. 2d) by altering the properties of the chamber surface to resemble that of the extracellular matrix (ECM). Proteins or enzymes can be immobilized onto surfaces to manipulate cell growth. Topographical features can also be incorporated to further mimic the ECM environment.

Surface Chemistry. As the size of devices decreases, their surface/volume ratio increases, making surface properties increasingly important in defining the performances of smaller bioreactors because cell function is intimately linked to the properties of the surfaces to which cells attach. Depending on the application, it is necessary to promote specific adsorption of proteins that mediate cell attachment and growth onto surfaces, or to prevent adsorption of proteins and cells. Because cellular receptors that bind to surfaces are nanoscale in size, it is also important to be able to pattern adhesive surfaces with resolutions from microns to nanometers (2,49–51). Surface properties are also important for enzyme-based microbioreactors because surface properties alter enzyme activity and stability. Enzymes are commonly immobilized through physical adsorption or covalent binding onto high surface area materials, such as carbon, silica, and polymers. Use of immobilized enzymes is often favored over free enzymes because of reduction in enzyme costs, ease of recovery, stability, and ability to be incorporated into microsystems. Ratner's recent review covers important topics including surface modification of materials to prevent nonspecific protein adsorption, immobilizing functional groups on surface, and development of synthetic materials (52).

A useful model surface for studying biomaterials interactions is a self-assembled monolayer (SAM). Often formed on gold using alkane thiols, SAMS are highly ordered arrays of linear molecules that have one end attached to the surface of gold or other bulk materials and the other end exposed to the environment. A useful feature of SAMs is that their surface properties are determined predominantly by the very terminal functional group. By altering the nature of these terminal groups, SAMs can prevent protein and cell attachment or promote binding of specific ones (53). Other types of systems that are useful for controlling protein adsorption include covalent bonding, via silanes, to silica or metal. Micropatterns of biopolymers can also be generated using photolithography, which uses patterns of light to induce region-selective chemical reactions on a surface. A more recent technique is microcontact printing, which involves transfer of small molecules or proteins onto solid substrates from an elastomeric stamp (53). Microfluidic networks have also been used for protein patterning: elastomers with embedded channel features are used to direct small volumes of protein solutions into networks of channels to create protein patterns corresponding to the path of fluid flow (54–56). Surface patterned microfluidic channels can then be used directly as microbioreactors in which micropatterned cell culture can be performed.

Topography Control. Living organisms are not flat. The endothelial lining of blood vessels, for example, exhibit an irregular wave-like topography to prevent build-up of fatty deposits (57). Endothelial monolayers have been modeled as a wavy surface by computational methods to estimate the influence of the waviness on local flow forces (58). Muscle fibers form microscale ridges and grooves onto which myoblasts can attach and proliferate (59). Microfabricated topographical features, regardless of whether they mimic physiological microtopographies or not, can be used in microbioreactors to modulate adhesion, align or orient cells, and even affect cell growth and differentiation. Examples of topographical features that have been studied include single cliffs, grooves/ridges, spikes, hills, tunnels and tubes, fibers, cylinders, mesh, waves, and random roughness. Materials used for topographical controls include gold, silicon, carbon, inorganic compounds, such as silica, lithium niobate, silicon nitride, and polymers, such as polymethylmethacrylate, silicones, cellulose acetate, collagen, fibrin, and PDMS (51). Microtopography can also affect surface wetting and fluidic flow patterns. Even for a simple groove structure, variations of the aspect ratio and contact angle of the underlying substrate materials can dramatically alter the morphology of liquid droplets contacting the surface (60).

Sterilization. Similar to larger scale bioreactors, microbioreactors and their solutions and gases can be sterilized using heat, chemicals, radiation, or through the filtration of agents. The small reaction volumes of microbioreactors provide an advantage in terms of sterilization, because for a given concentration, the total number of cells, spores, or other contaminants depends on the volume. Then, assuming that the "death rate" of the contaminant is independent of reactor size or contaminant numbers, the sterilization time needed to extinct the contaminants will decrease with decreasing reactor size. In this respect, microbioreactors are more time efficient in batch sterilization than macroscopic ones. Ultraviolet (UV) sterilization is particularly convenient for transparent microbioreactors, such as those fabricated in PDMS. When pH sensors or dissolved oxygen sensors are packaged into the microbioreactor, autoclave and UV radiation may not be feasible, and alternative methods may need to be used. Flowing 70% ethanol through the chambers/channels and subsequently drying is also sufficient for many microbioreactor applications.

Other Considerations. Phototrophic microorganisms consume light energy to survive; they can grow in simple and inexpensive nutrient media. The light requirements of phototrophic microorganisms, however, impose other significant constraints on photobioreactor designs (61). These constraints are due to an exponential attenuation of the light flow passing through an optically absorbing medium. Microbioreactors, given their large surface/volume ratios, are promising for photobioreactor applications. Cultivation of phototrophic microorganisms in optimized photobioreactors would increase the product yield several folds by maintaining the culture under appropriate conditions.

Ultrasound irradiation can change both the structure and function of biologic molecules such as proteins and deoxyribonucleic acid (DNA) (62). At mild intensity, ultrasound irradiation can increase the activity of free enzymes (63). For example, significant enhancement of reaction rate can be achieved when exposing subtilisin power to ultrasound irradiation (64). Low energy ultrasound wave irradiation can also optimize the efficiency in ethanol production by yeast from mixed office waste paper in bioreactors (65). A variety of microscale ultrasound systems have been reported. Advanced microbioreactor systems with ultrasound components may be developed to make bioreactors more efficient.

MICROBIOREACTOR COMPONENTS

Components For Heat Transfer

External heating elements can be incorporated into microbioreactors to control temperatures. External controls of temperature include the use of standard cell culture incubators that can accommodate the whole microbioreactor (66), heating tapes, and water-to-water heat exchangers (67). Internal, embedded heaters can be comprised of materials such as thin platinum films (68) or optically transparent indium–tin oxide (ITO) (69). Temperature measurements can also be made on chip using metallic, semiconducting, or optical materials.

An important application where heaters are crucial and microbioreactors have an advantage is the polymerase chain reaction (PCR), a temperature-controlled and enzyme-mediated DNA amplification technology (70). Polymerase chain reaction requires multiple cycles of high, low, and medium temperatures to separate DNA strands, anneal the primer to the template DNA, and make complementary copies of the template DNA. Since microsystems usually have high heat conductivity and low heat capacity, the time for raising and lowering the temperature is shortened and time will be saved when cyclic temperature fluctuations during the PCR reaction is necessary. Northrup et al. (71) integrated microfabricated polysilicon heaters into a micromachined silicon reaction chamber. Schneegass et al. (68) used a thermocycler chip with integrated thin platinum film heaters and sensors for temperature. Kopp et al. (72) used external copper blocks and heating cartridges with the surface temperature monitored by a platinum thin-film resistor. Burns and co-workers (73) developed a very simple and elegant PCR device that utilizes Rayleigh–Benard convection—a steady, buoyancy-driven circulatory flow that occurs between two surfaces, one on top and one on the bottom, maintained at two fixed temperatures—to perform temperature cycling.

Components For Aeration

Electrolytic gas generation provides a compact, scalable approach for gas delivery to microbioreactors (19). In brief, a pair of interdigitated Ti/Pt electrodes hydrolyzes electrolyte to generate oxygen gases at the narrow end of a gradually widened hydrophilic microchannel. The oxygen bubbles move along the conical microchannels and transfer into culture medium because of the positive pressure built up during gas generation and the different surface tension forces at the front and back of the bubbles formed in the gradually widening channel. The rate of oxygen generation can be precisely controlled by pulse width modulation of the electrode potential.

The smaller the bioreactor, the more crucial it is to avoid bubble formation to prevent blockage of microchannels. For such systems, it is advantageous to use gas permeable membranes that allow diffusion of gases through it without introduction of bubbles. Because the surface/volume ratio is large in microsystems, oxygen transfer through gas-permeable membranes is often sufficient to ensure adequate oxygenation for biomass production. A straightforward method to fabricate gas-permeable membranes is to spincoat PDMS prepolymer onto silanized silicon wafers, cure and harden to generate a thin membrane, then peel it off (20,21).

Components For Fluid Control

Valves. Valves can be categorized into active and passive valves. A passive valve is a flow-dependent obstruction that functions without any external actuation. Passive valves are mostly unidirectional. In an active valve, fluid flow is directed by active actuation (74). The advantage that active valves have over passive valves is the degree of control one has over the timing, rate, and direction of fluid flow. This type of control is necessary to make real-time adjustments and for feedback control of microbioreactors. Although a large variety of valves have been reported, it is still a challenge to integrate multiple valves into a functional bioreactor system. Difficulties arise because many valves are incompatible with other components to be integrated, or because the valves require large external systems for actuation.

Passive valves have been used for restricting flow to one direction, removing air from liquid, or making flows stop at select channel regions. Although the level of control is lower compared to active valves, passive valves have the advantages of having few or no moving parts, less complexity, easy fabrication, and less chance to break due to fatigue (75). Recent approaches for passive control valves involve the use of hydrophobic materials, surface patterning, and changing channel fluid resistance (by changing channel geometry) (11). Passively moving micropiston valves have also been fabricated *in situ* inside microchannels using laser polymerization of a nonstick polymer (76).

Most conventional active microvalves couple a flexible diaphragm (77,78) to, thermopneumatic (79), piezoelectric (80), electrostatic, electromagnetic (81), bimetallic, or other types of actuators. The scaling of these actuation forces to the microscale, however, is often unfavorable and requires macroscale external actuators for operation. An interesting alternative to active valves is the use of autonomously regulated valves made of hydrogels that swell in response to pH or other specific chemical or thermal environment (11,82). The volume changes of the hydrogel can valve or obstruct fluid flow directly, or indirectly, through deformation of a thin PDMS membrane. The PDMS, when deformed, partially occludes an orifice to regulate the feedback stream of compensating buffer solution (17). By altering their chemistry, hydrogels can also valve in response to the changes of temperature, light, electric fields, carbohydrates, or antigens. Ehrick et al. (83) incorporated genetically engineered proteins within hydrogels that swell in response to various ligands as potential valves for microfluidic channels. Other types of valves include the use of commercially available Braille displays (84) or a pneumatic valve system to deform flexible microchannels (85).

Pumps. A micropump should ideally be able to pump a wide range of fluids and gases, be self-priming, and be programmable. Ideal micropumps are still lacking; thus, many microfluidic applications use macroscopic pumps, such as syringe pumps. Micropumps can be classified into two main types: mechanical pumps and nonmechanical pumps (71). Mechanical pumps use electromagnetic, piezoelectric, pneumatic, shape memory, electrostatic, thermopneumatic, or thermomechanic components to deliver fluid. Mechanical pumps provide higher control over average flow rates, but the flow is often pulsed and the fabrication relatively complex (53,85). Many types of nonmechanical micropumps have also been successfully developed. The flow from nonmechanical pumps is usually pulse-free with a wide range of flow rates at low pressures and the fabrication is often less complex compared to mechanical pumps. An electrokinetic pump that utilizes an electric field for pumping conductive fluids by electrophoresis or electroosmotic flow (EOF) is the most common method to control flow in microfluidic systems (71). Electrokinetic pumps have the advantages of direct control, fast response, and simplicity. However, the substrate material, joule heating effect, and microchannel charge have to be considered. Other types of nonmechanical pumps include electrohydrodynamic pumps that use electrostatic forces acting on dielectric fluids, phase-transfer pumps that use pressure gradient between gas and liquid phases, electrowetting fluid actuation systems that use interfacial forces, electrochemical pumps that use the pressure of gas bubbles generated by electrolysis of water (71), magnetohydrodynamic pumps (86), capillary force, gravity-driven pumps (87), pneumatic pumps (85), and pumps that use action of piezoelectric pin arrays in refreshable Braille displays (84).

Multiple Laminar Streams. For most flows in small channels, viscous forces dominate; thus flow is laminar and lacks turbulence. When two or more streams pass through microchannels, they flow in parallel as if they are separated by physical boundaries. Laminar flow is a challenge for mixing, but a useful phenomenon for microscale fluid patterning (Fig. 3d) (55,88). By taking advantage of diffusive mixing (but not turbulent mixing) and laminar flows, spatiotemporally defined gradients can be generated and have been used to study chemotaxis (89,90). Multiple laminar flows have also been used for developing microfluidic assay systems [i.e., T-sensor (91)], and studying subcellular processes when the interfaces of the laminar streams are positioned over a single cell (92).

Mixers. Mixing is challenging in microfluidic systems because laminar flows preclude turbulent mixing (Fig. 3d). Microscale mixers, therefore, generally use elongational flows or laminar shears to increase interfacial areas between different fluids and mixing by diffusion. Distributive mixing physically splits the fluid streams into smaller segments and redistributes them to reduce the striation thickness. Passive and active mixers have been developed for microfluidic systems, including laminating mixers (93), rotary mixers (94), mixing based on out of phase forward and backward pumping of different liquids (84), plume mixers (nozzle arrays), chaotic advection mixers (9,95), movement of liquid plugs (96), and an electroosmotically driven micromixer that uses multiple intersecting channels to enhance lateral transport (97). Thorough reviews on micromixers have been given by Hessel et al. (98) and Nguyen and Wu (99).

Components For Mechanical Stimulation

Shear. Methods commonly used to impart shear stress on cells include cone viscometers, parallel plates, and capillary tube flow chambers (100,101). Gradients of shear stress can also be generated in a curved D-shape microchannel (102). Flow-induced cytoskeleton rearrangements were shown to depend on the geometry of the channel (D-shape channel versus flat surfaces, representing experience in microcirculation and large veins, respectively) and the presence of inflammatory drugs (103).

Stretch. When subjecting cultured cells to mechanical stretch, proper design and application of a strain device are required to provide a well-defined and reproducible strain field to study mechanotransduction. Information from such *in vitro* models, which facilitate systematic variations in mechanical conditions and allow rapid analyses, would yield tremendous insights into mechanical parameters that may be important *in vivo* (104). Biaxial cell strain devices have been used to strain lung cells (105,106) by repeated mechanical deformations of a membrane on which cells are attached. Strain could also be applied using a magnetic force (107), or via uniaxial cyclical stretch (108).

Components For Separation And Purification

Lysing. Cell contents are separated from their surrounding environment by a cell membrane. The membrane and an underlying cytoskeletal network provide mechanical strength to the cell and preserve its integrity. The first step in many analyses or isolation of cell contents is to disrupt the cell membrane. There are a variety of mechanical (homogenization, milling, ultrasonic disruption, and blenders) and nonmechanical (detergent, organic solvent, osmotic shock, enzymatic permeabilization, electrical discharge, heating, and pressure cycling) methods for disrupting cell membranes on the macroscopic scale. Demonstration of cell lysis on the microscopic scale inside microfluidic devices, however, has mostly been with nonmechanical methods that use detergents (96,109), electrical discharges, or lasers (110). Miniaturized cell electrolysis devices can work with small amounts of cells and reduce the amount of purification compared to other protocols (111,112). For example, Waters et al. (113) developed a microchip that is capable of performing *Escherichia coli* lysis, PCR amplification, and eletrophoretic analysis on a single device.

Separation. Cell separation techniques are fundamental to clinical diagnosis, therapy, and biotechnological production (114). For example, it is crucial to have purified cells before proliferation and production of cellular products. Current approaches include optical tweezers (115), centrifugation (116,117), filtration

(109,116,118), fluorescence-based cell sorting (FACS) (119,120) or magnetically activated cell sorting (MACS) (121), electric field-based manipulations and separations (114,122–124), and cell-motility based sorting (125,126). Microtechnology opens new opportunities in cell and biomolecule sorting that take advantage of laminar flow behaviors (125), electrical field properties (127), or other microscale phenomena. It also provides the opportunity to combine multiple modes of separation into an integrated system. Researchers have used physiological fluid mechanical phenomenon observed in blood microcirculation (plasma skimming and leukocyte margination) to filter leukocytes from whole blood (128). Petersson et al. (129) combined acoustic wave forces and laminar flow to continuously sort erythrocytes from lipid particles in the whole blood. Because these two components were different in density and responsiveness to pressure, erythrocytes were enriched at the pressure node (along the center of the channel) and lipid particles were gathered at the pressure antinodes (along the side walls). Finally, the erythrocytes and lipid microemboli separated into different branches at the end of the main channel. Because of the variety of different properties by which cells of interest need to be sorted, it is important to develop multiple modes of cell sorting. Reviews about microfluidic cell analysis and sorting devices can be found elsewhere (130,131).

Filtration. Microfabrication techniques have been developed to integrate filters and membranes inside microbioreactors. Zhao et al. (132) and Hisamoto et al. (133) successfully produced semipermeable nylon membranes inside microfluidic channels, by taking advantage of laminar flow so that a polymerization reaction would occur at the liquid interface of the two flows and produce a thin membrane in predetermined areas of a microfluidic device.

Components For Monitoring And Control

A key requirement for microbioreactors is the ability to measure parameters, such as temperature, dissolved oxygen, pH, and flow rate. For systems involving cells, mass balances are even more complex than with enzyme bioreactors because of the larger number of products and byproducts produced and the complex responses of cells to the changes in material concentrations. In small volumes, it is difficult or impossible to use standard, macroscopic, industrial probes. Miniaturized sensors must be developed (17). Sensors that do not consume the analytes are also preferred to avoid depletion and also because sample extraction is hard to achieve in closed and compact microsystems without disrupting the devices.

Optical, electrochemical, and thin-film solid-state conductivity are the three main categories of microsensing schemes. Many of the most useful microdetection schemes are based on optical measurements such as fluorescence intensity (134), fluorescence lifetime, chemiluminescence (135), and bioluminescence (136). Optical sensing is convenient because molecular or nanoscale "sensors" are simple to introduce inside microchannels and readout can be detected from a distance without direct external contacts. In addition, many optical probes can sense without consuming oxygen or perturbing pH. Nonperturbing sensors are important for maintaining a constant microenvironment because the quantities of chemicals are small in microbioreactors. Bioluminescence and chemiluminescence are sensitive with the detection limits down to 10^{-18}–10^{-21} mol, which offers great advantage over other spectroscopic-based detection mechanisms. Laser-induced fluorescence detections in microsystems have been reviewed by Johnson and Landers (137). Other optical detection methods for microfluidic systems have been reviewed by Mogensen et al. (138).

Sol–gel-based probes encapsulated by biologically localized embedding (PEBBLEs) allow real-time measurement of molecular oxygen, pH, and ions inside and around living cells (139). Kostov et al. (17) used an optical sensing system integrated with semiconductor light sources and detectors to perform continuous measurements of pH, optical density, and dissolved oxygen in miniature bioreactors. A drawback of optical sensing is the need for light sources, lenses for focusing, and detectors, which are more challenging to miniaturize compared to the sensor probes themselves. A useful solution is to use optical fiber-based systems, which allows decoupling of the probes from the light sources and detectors, and enables detection at sites inaccessible by conventional spectroscopic sensors (140). Compared with silicon micromachining techniques, the fabrication and integration of polymeric optical elements (waveguides, lenses and fiber-to-waveguide couplers) with microfluidic channels are fast and simple (141). An oxygen-sensitive fluorescent dye has been developed to monitor dissolved oxygen levels in a system. This provides advantages over electrochemically based sensors [for a review, see Ref. 142] due to their size, ease of fabrication, and sensitivity (141).

Electronic microsensor is another category that contains a large number of useful biochemical detectors (143). Electronic microsensors usually require direct hard wiring of sensors to a readout system but can be easier to multiplex and are often smaller overall compared to optical systems. For example, Walther et al. (144) integrated a pH-ISFET (ion-sensitive field-effect transistor), a temperature-sensitive diode, and a thin-film platinum redox electrode on a single chip. The chip is mounted on a carrier and inserted into the chamber to monitor various biological parameters such as pH and redox potential. Brown and coworkers developed polymer membrane-based solid-state sensors to measure pH, and ions (145).

Microfabricated ultrasensitive nanocalorimeters can measure heat generation to monitor cell metabolic activity in response to agonist and antagonist using as few as 10 cells and without prior knowledge of the mode of action of these drugs. This measurement is noninvasive and quantitative and is envisioned to be useful for pharmaceutical companies to find drug candidate (146).

Li et al. (147) developed a microfabricated acoustic wave sensor to measure the stiffness of a single cell. This sensor is promising for drug screening and toxicology studies. For example, the acoustic wave sensor is envisioned to measure the rigidity of a heart cell to distinguish the effect of positive and negative ionotropic drugs. Positive ionotropic drugs are useful for treating congestive heart failure and negative ionotrpic drugs for hypertension. The acoustic wave sensor is also interesting for single cell muscle physiology.

Cells themselves can be used to sense and amplify biological signals. Several groups have developed histamine sensors by integrating cells on microfluidic devices (148,149). This chip-based detector caters to the need for a simple, rapid, and safe method for allergy identification. Cells can be engineered to have a variety of biological recognition events coupled to reporter genes to specifically sense analytes of interest (150).

Some new exciting prospects for future biosensors are in the area of nanotechnology. For example, quantum dots (151) and nanoscale PEBBLES (139) are extending the limits of sensitivity, stability, biocompatibility, and flexibility in optical sensing. Quantum dots are small semiconductor nanocrystals (on the order of nanometers to a few micrometers). Fluorescent quantum dots are able to detect biological species by fluorescing only when coming in contact with viable cells, making them useful probes for many types of labeling studies. These quantum dots are photobleached very slowly and can be manufactured to emit a wide range of wavelengths. They can be used in cell biology for the labeling of cellular structures, tracking the fate of individual cells, or as contrast agents (152). Nanowire-based sensors with their small size and higher sensitivity would also be ideal for integration into microbioreactors (153). Many microsensors have been developed and are ready for integration into microbioreactors.

Fabrication

A variety of methods and types of substrates are available for microfabrication. The most traditional and widely used method of microfabrication is photolithography. Originally developed for the microelectronics industry, photolithography is precise, highly reproducible, and capable of mass production. Photolithography uses patterns of UV light coming through a photomask to area-selectively induce chemical reactions in a polymeric, light-sensitive photoresist coated onto a semiconductor substrate. During development, the light exposed regions of the photoresist are selectively removed or selectively left behind generating micropatterned photoresist structure. The exposed areas of the substrate are then chemically etched to provide features of various depths and shapes.

Due to the high equipment costs involved in photolithography and because silicon and glass substrates used in electronic and mechanical devices are not necessarily the best materials for biological applications, alternative types of microfabrication have been developed. A cost-effective and experimentally straightforward method called soft lithography has been particularly useful for biological applications (2,53). Soft lithography uses elastomeric materials, such as PDMS to create microstructures, and to pattern and manipulate surfaces. The process involves casting PDMS against a photolithographically defined master mold to yield a polymeric replica. The PDMS replica is then sealed against another material to form channels and reservoirs. Alternatively, the replica can be used in a technique called microcontact printing, where the PDMS mold is used as a stamp to transfer protein or molecular ink to a substrate. The PDMS has several properties, which make it useful for biological applications: (1) biocompatibility allows cell culture on and inside PDMS structures, (2) optical transparency allows optical inspection and sensing, (3) gas permeability allows long-term growth of cells without depletion of oxygen, (4) flexibility allows cell cultured on PDMS to be mechanically stretched (2,50,53).

Other polymer based microfabrication techniques include hot embossing, injection molding, and laser ablation. A hot embossing technique called nano-imprint lithography developed by Chou et al. (154) has the ability to fabricate sub-10 nm nanometer features. This process creates nanostructures in a resist by deforming the resist shape with embossing (155). Moriguchi et al. (156) developed a unique photothermal microfabrication technique, where agar microchamber arrays with living cells inside them can be remolded *in situ* during cell cultivation.

Integration

The development of a microbioreactor requires assembling multiple functional units (for electronic, mechanical, biological, and chemical processing) into a compact device. Integration and packaging poses a whole new challenge on top of the challenges to develop individual components. With macroscopic bioreactors, it is relatively straightforward to connect different components together with little or no worries of space organization. As one scales down, the placement of components must be performed strategically as the room around the reactor chamber decreases dramatically (because volume scales as length cubed, a 10 time reduction in linear dimensions, e.g., will lead to a 1000 time reduction in the available space). There are also technical challenges to fabricate microscale fittings and connectors, or to join two components via connectors even if they could be fabricated. A variety of processes used to build integrated circuits and microelectromechanical systems have played a major role in fabricating integrated microfluidic systems and microbioreactors. These technologies, known collectively as micromachinging, selectively etch away or deposit structural layers on silicone wafers.

Recent efforts to reduce costs, enhance material biocompatibility, and needs for diverse chemical and mechanical properties have also led to the use of a wide variety of polymeric materials in microfabricated devices. Integration, unless planned carefully, can lead to material incompatibilities in fabrication or operation. Notable accomplishments of integrated microfluidic systems include DNA analysis chips (78,157), pneumatically driven microfluidic cell sorters and protein recrystallization chips (77,119,158,159), portable cell-based biosensor systems (160), microfermentors with integrated sensors (17,20,21), and a computerized microfluidic cell culture system actuated using the pins of a refreshable Braille display (84). Integrating fluid control components for cell-culture microbioreactors is rapidly progressing, but still underdeveloped. Interested readers are referred to recent review articles on integrated microfluidic devices (161).

SPECIFIC EXAMPLES OF MICROBIOREACTORS

This section presents select examples of microbioreactors. The examples are not exhaustive, but are meant to show representative examples in each of the following four categories: (1) microbioreactors that are used to optimize bioproduction, (2) microbioreactors that provide cells with physiological microenvironments to more accurately predict physiological drug kinetics and toxicity, (3) microbioreactors that are used to develop cell-based therapies, and (4) microbioreactors for mechanistic studies. Because the field is still young, the devices are relatively simple and many are still prototypes rather than refined products ready for real world applications. The rapid advances, however, promise an increasing role of microbioreactors in the clinic, laboratory, and at home or other points of need.

Microbioreactors For Optimizing Production Conditions

With the development of microtechnologies, more and more bioprocess optimization is performed in small volume bioreactors with integrated detection systems. For example, Rao and co-workers (17) have demonstrated parallel fermentations of *E. coli*. in a milliliter-size microbioreactor. A smaller, microliter-size fermentor has been developed recently by the Jensen's group (20). The performances of these microbioreactors are comparable to traditional liter-size fermentors: Measurements of pH, dissolved oxygen (DO) and optical density (OD) of biomass in these microbioreactors have similar profiles as those in benchtop fermentors. Similarities in cellular metabolism and growth show the potential of using microbioreactors for bioprocess optimization. Arrays of microfermentors with integrated sensors and actuators are envisioned to drastically reduce the cost and time for developing new bioprocesses.

Microbioreactors For Toxicological And Drug Testing

Drugs need to be tested for toxicity and efficacy before administration to humans. Accurate prediction, however, is a challenge because most current drug tests are performed only on human cells or animals. Animal tests are expensive and time consuming, and efficacy of a drug obtained from an animal surrogate study can still be difficult to extrapolate to humans (162). Even when one uses human cells for analysis, the result may be totally different from what occurs physiologically because cells cultured in flasks or dishes experience a totally different microenvironment. Therefore, a microscale human surrogate with microcirculatory systems, three-dimensional (3D) tissue organizations, and appropriate cell–cell and tissue–tissue interactions would be beneficial in predicting human responses to drug treatment more precisely. Below are two notable examples of such efforts.

Bioartificial Livers. There are two major applications for which artificial livers are developed: one is to replace organ functions in patients with liver failure, and the other is to perform toxicology testing of drug candidates. Organ replacement functions require large bioartificial livers (BALs), whereas the toxicology studies would benefit from microscale BALs capable of conducting high throughput analyses. Microscale BALs are promising as convenient and low cost *in vitro* models for screening drug toxicities particularly in light of the fact that approximately one-half of all drug toxicities involve the liver.

Liver failure is the seventh leading cause of death by disease in the United States. About 26, 000 people died each year because of liver failure. Transplantation is limited by the supply of donor organs and the cost of the surgery (163). Extracorporeal BAL devices have been proposed as substitutes for transplantation. Macroscopic BALs have been tested in clinical trials (164). In an attempt to maximize the efficacy of the BALs, and to develop *in vitro* liver systems for biological and toxicological studies, several groups have microfabricated liver cell culture systems (165).

Microbioreactors for liver cell cultures have several configurations, ranging from flat-plate (163) or matrix-sandwiched monolayer designs (166) to 3D perfusion cultures (165). A flat-plate microbioreactor with an oxygen permeable membrane was shown to support viability and synthetic functions of hepatocytes cocultured with 3T3-J2 fibroblasts (163). This microchannel bioreactor was also functional when connected extracorporeally to a rat (162). The results indicate that this device can potentially be used as a liver support device and for the eventual scale-up to clinical devices. Compared with other configurations, monolayer designs excel in mass transfer, easy fabrication, and easy optical analysis of cells (165), although cells might be damaged by exposure to shear stress (167). Griffith and co-workers (165) developed an array of microbioreactors (with each channel $300 \times 300 \times 230$ μm, $L \times W \times H$ in dimension) that support 3D culture of liver cells by perfusion. Liver cells cultured in this device showed viable tissue structures and tight junctions, glycogen storage, and bile canaliculi. Membrane-based 3D perfusion hepatocyte culture systems have also been developed (66).

Micro CCA. It is important to integrate cells from different tissues together to simulate physiological drug metabolism. Shuler and co-workers developed Cell Culture Analogs (CCAs) that combine mathematical pharmacokinetics models with cell culture-based experimental studies to mimic human responses. The CCAs have compartmentalized cell cultures representing different tissues. Interconnections between these compartments allow circulation of media and metabolites. Since the CCAs mimic the time-dependent exposure and the metabolic interaction between multiple types of tissues and cells, predictions from the CCAs may be more accurate compared to existing *in vitro* models. Shuler and co-workers proved the concept of CCA with a macroscopic three-compartment (liver, lung, and other tissues) system and showed the feasibility and potential usefulness of such devices in testing naphthalene toxicity (168,169). MicroCCAs with three (liver, lung, and other tissues) and four (liver, lung, fat, and other tissues) compartments have also been reported (170,171).

Microbioreactors For Therapeutical Applications

Microfluidic Systems As Assisted Reproductive Technologies. Microdevices provide unique platforms for artificial reproduction and may ultimately increase the efficiency, safety, and cost-effectiveness of *in vitro* fertilization procedures. Currently, many embryo experiments are performed in macroscopic culture dishes, where human or animal oocytes (eggs) and embryos are manipulated manually. Use of pipettes for cellular manipulations is labor intensive and low in accuracy, reproducibility, and efficiency. In addition, the practice of transferring embryos from one type of media into another is abrupt and may shock the embryo due to the sudden change of environment (172–174); inside the female tract, the supply of nutrients, growth factors, and hormones changes gradually as the embryo development progresses. An alternative to manual pipetting is the use of microfluidic channels. Microfluidics is ideal for use in artificial reproduction, because it is a procedure that occurs physiologically inside small tubes and ducts, the size and numbers of cells (oocytes, sperms) required match well with dimensions of microsystems.

Beebe et al. (175) demonstrated manipulations of embryos and oocytes within microfluidic channels. The microfluidic systems can transport single mouse embryos through a channel network (176), remove the zona pellucida by chemical treatment (177), remove cumulus cells from oocytes via mechanical suction (178), and culture embryos in static or dynamic fluid environments (179). In some cases, embryos cultured in microfluidic channels develop faster compared to embryo grown in culture dishes and with growth kinetics that are closer to what is observed *in vivo*.

Cho et al. (125) developed a Microscale Integrated Sperm Sorter (MISS) that isolates motile sperms based on the ability of the motile sperms, but not the nonmotile ones, to cross-laminar flow streamlines. The device allows small volume samples that are difficult to handle with conventional sperm-sorting techniques to be sorted efficiently using a mild biomimetic sorting mechanism. The MISS integrates power source, sample injection ports, and sorting channel into one disposable polymer device making the device potentially useful not only clinically, but also as an at-home male infertility test (180).

Implantable Microcapsules. Microbioreactors that produce and release natural or recombinant bioagents are useful for delivering therapeutic agents *in vivo* to enhance metabolic function (181) or treat neurological disorders (182) and cancers (183,184). Mammalian cells, plant cells, microorganisms, enzymes, and biochemical compounds have been encapsulated in a variety of semipermeable containers mainly made of synthetic and natural hydrogels (e.g., poly(vinyl alchohol), poly(hydroxyl ethyl methacrylate), calcium alginate κ-carrageenan, chitosan, collagen, and gelatin). Here we focus on the application of microencapsulated mammalian cells. The outer membranes of the microspheres encapsulating the cells serve as selective barriers that allow exchange of nutrients, wastes, and therapeutic agents but block the passage of encapsulated cells as well as macromolecular components of the immune system. This approach has been effective in delivering genetically engineered cells that secrete growth hormone to partially correct growth retardation (185), recombinant human bone morphogenetic protein-2 (rhBMP-2) to induce bone formation and regeneration (186), interleukin-2 to delay tumor progression and prolong survival (187), coagulation factor IX for treatment of hemophilia B (188), dopamine to treat Parkinson's disease (182), insulin to maintain blood glucose level (189), and analgesic substances to relief pain (190). Most of these works are aided by using murine and canine models. Some of the most advanced systems are in early clinical trials. Biocompatibility (191,192) mechanical stability of the capsule material (193–195), efficacy (196), safety (197), and cost are still under evaluation and optimization. Reviews of therapeutic uses of microencapsulated genetically engineered cells can be found elsewhere (198).

Microbioreactors For Understanding Biological Responses

Use Of Multiple Laminar Streams To Study Subcellular Biology And Chemotaxis. Physiological cell environments are heterogeneous with local production and consumption of key growth factors, nutrients, and signaling molecules. Microfluidic systems are useful for mimicking such micropatterns of chemicals around cells. A particularly simple and useful method is to take advantage of small channel dimensions to generate multiple laminar streams that flow in parallel and adjacent to each other inside the same microchannel with minimum mixing (Fig. 3d). Such techniques can be even used to treat different parts of single living cells with different small molecular drugs, proteins, and small particles such as low density lipoprotein (LDL) (88).

Cancer and normal cells receive similar local stimuli inside the body, but behave totally differently. A critical question is why these differences arise. Taking advantage of multiple laminar flows to perform subcellular epidermal growth factor (EGF) stimulation, Sawano et al. (92) revealed differences in signal propagation between carcinoma and normal cells in response to local EGF stimulation.

Many cells direct their motion in response to chemical gradients. This phenomenon, called chemotaxis, protects microorganisms by allowing them to move toward more favorable conditions. In mammals, chemotaxis is important for guiding cell migration during development, embryogenesis, cancer metastasis, and inflammation. A challenge for analyzing chemotaxis is the lack of methods to generate well-defined chemical gradients that are stable and do not change with time. This difficulty arises due to the diffusivity of molecules and resulting changes of concentration profiles over time. Jeon et al. (90) demonstrated the use of microfluidic systems with branched networks of channels to generate stable gradients of interleukin-8 (IL-8) with linear, parabolic, and periodic concentration profiles. The position and shape of the concentration gradients were controlled by adjusting the flow rates and the positions to which reagent of interest were added into the channel network (89,199). The well-defined gradients allowed straightforward quantification of chemotaxis coefficients as well as to observe complex migration behaviors of leukocytes in response to different concentration gradient profiles.

Studies Of Vascular Diseases In Microfluidic Channels. The mechanical forces that accompany blood flow and pressure fluctuation influence vascular cellular biology and pathology in many ways. As the blood flow along a vessel, the viscous drag forces constantly expose endothelial cells (ECs) to shear stress. The pulsatility of the blood flow also induces a periodic change of circumferential strain on ECs and their underlying smooth muscle cells (SMCs). These mechanical forces have been found to cause important biological changes in endothelial cell morphology and function, such as alignment and elongation (30,200), low density lipoprotein uptake (201), tissue plasminogen activator synthesis and secretion (202), and proliferation (67). There has also been studies about the combined effect of shear stress and cyclic strain on ECs (203–205) and SMCs (206–208). While much has been revealed about the alterations in EC function induced by mechanical stresses, relatively little is known about the mechanism mechanical signaling.

Capillary-size microfluidic channels were used to model malaria, a potentially vital disease caused by loss of deformability of red blood cells due to *P. falciparum* parasites infection. Erythrocytes exhibited increased rigidity and decreased deformability with the progression of the disease, as demonstrated by their increased difficulties to transverse through 2 to 8 μm wide PDMS channels. This type of microfluidic system may potentially be useful to screen antimalaria drugs (209).

CONCLUSION AND FUTURE PROSPECTS

The convergence of bioreactors with advances in microtechnology is starting an exciting revolution in medicine. With microfabrication technologies, many copies of a device can be generated and operated with quick procedures and reduced cost. The ability to perform parallel assays allows high throughput optimization of bioprocess conditions, opening the way for cost-efficient biopharmaceuticals production.

Humans and other living organisms are inherently microscopic in their essence, being comprised of networks of microscopic reactors (i.e., the cells), and interconnected by microfluidic vasculatures. Efforts to miniaturize bioreactors would lead to not only the development of smaller pharmaceutical production and screening systems, but also the construction of more physiological *in vitro* cell culture systems where the ultimate goal is to develop microbioreactors as animal or human surrogates. Even for cullture of single cell types, the ability of microfluidic systems to simulate physiological microenvironments is useful for revealing disease mechanisms, drug testing, use as biosensors, and single-cell-based clinical procedures such as *in vitro* fertilization. More complex microbioreactor systems with multiple cell types are being developed for toxicology studies. So-called animals-on-a-chip or minihumans provides exciting prospects for efficient drug discovery and personalized medicine.

Current state-of-the-art microbioreactors are still relatively simple with few components and limited sensing and control. Many are highly specialized and nonroutine in their use. The overall footprints of the systems are also often still macroscopic. Current advances in micro- and nanotechnologies as well as in medicine, however, promise rapid improvements in performance, accessibility, and sophistication of microbioreactors. Current trends point to a future where the gap between manmade devices and living organisms will narrow and applications of microbioreactors to medicine will grow.

ACKNOWLEDGMENTS

We thank the Whitaker Foundation, National Science Foundation (BES-0238625, DMI-0403603), National Institutes of Health (EB00379-01, HD049607-01), NASA (NNC04AA21A), and the U.S. Army Research Laboratory and the U.S. Army Research Office under contract/grant number DAAD19-03-1-0168 for financial support. We thank G. D. Smith, S. C. Chang, H. Chen, P.Y. Choi, B. H. Chueh, T. Kable, E. O. Kass, K. Ke, J.H. Kim, S. M. K. Lau, E.C. Lee, W.H. Lee, S.R. Mandal, J.A. Miller, Y. Murgha, S.O. Charoen, A.B. Ozel, S.J. Segvich, P.D. Settimi, D. Sud, B. Teply, M. Zhang, C. Zhong for assistance.

BIBLIOGRAPHY

1. Walker GM, Zeringue HC, Beebe DJ. Microenvironment design considerations for cellular scale studies. Lab Chip 2004;4:91–97.
2. Shim J, et al. Micro- and nanotechnologies for studying cellular function. Curr Top Med Chem 2003;3:687–703.
3. Bhadriraju K, Chen CS. Engineering cellular microenvironments to cell-based drug testing improve. Drug Discov Today 2002;7:612–620.
4. Desai TA, et al. Microfabricated immunoisolating biocapsules. Biotechnol Bioeng 1998;57:118–120.
5. Krenkova J, Foret F. Immobilized microfluidic enzymatic reactors. Electrophoresis 2004;25:3550–3563.
6. Madou MJ. Fundamentals of Microfabrication. Boca Raton (FL): CRC Press; 2002.
7. Tabeling P, et al. Chaotic mixing in cross-channel micromixers. Philos Trans R Soc London Ser A-Math Phys Eng Sci 2004;362:987–1000.
8. Stremler MA, Haselton FR, Aref H. Designing for chaos: Applications of chaotic advection at the microscale. Philos. Trans R Soc Lond Ser A-Math Phys Eng Sci 2004;362:1019–1036.
9. Stroock AD, et al. Chaotic mixer for microchannels. Science 2002;295:647–651.
10. Squires TM, Bazant MZ. Induced-charge electro-osmosis. J Fluid Mech 2004;509:217–252.
11. Ehrfeld W, Hessel V, Lowe H. Microreactors: New Technology for Modern Chemistry. Chichester : Wiley-VCH; 2000.
12. McDuffie NG. Bioreactor Design Fundamentals. Boston: Butterworth-Heinemann; 1991.
13. Roels JA. Macroscopic Thermodynamics and the Description of Growth and Product Formation in Microorganisms. Washington, (D.C.): American Chemical Society; 1983.
14. Roels JA. Energetics and Kinetics in Biotechnology. Amsterdam: Elsevier Biomedical Press; 1983.

15. Ezashi T, Das P, Roberts RM. Low O-2 tensions and the prevention of differentiation of hES cells. Proc Natl Acad Sci U S A 2005;102:4783–4788.
16. Csete M, et al. Oxygen-mediated regulation of skeletal muscle satellite cell proliferation and adipogenesis in culture. J Cell Physiol 2001;189:189–196.
17. Kostov Y, Harms P, Randers-Eichhorn L, Rao G. Low-cost microbioreactor for high-throughput bioprocessing. Biotechnol Bioeng 2001;72:346–352.
18. Blanch HW, Clark DS. Biochemical Engineering. New York: Marcel Dekker; 1997.
19. Maharbiz MM, et al. A microfabricated electrochemical oxygen generator for high-density cell culture arrays. J Microelectromech Syst 2003;12:590–599.
20. Zanzotto A, et al. Membrane-aerated microbioreactor for high-throughput bioprocessing. Biotechnol Bioeng 2004; 87:243–254.
21. Maharbiz MM, Holtz WJ, Howe RT, Keasling JD. Microbioreactor arrays with parametric control for high-throughput experimentation. Biotechnol Bioeng 2004;85:376–381.
22. Irazogui G, Villarino A, Batista-Viera F, Brena BM. Generating favorable nano-environments for thermal and solvent stabilization of immobilized beta-galactosidase. Biotechnol Bioeng 2002;77:430–434.
23. Hara M, Adachi S, Higuchi A. Enhanced production of carcinoembryonic antigen by CW-2 cells cultured on polymeric membranes immobilized with extracellular matrix proteins. J Biomater Sci-Polym Ed 2003;14:139–155.
24. Keane JT, Ryan D, Gray PP. Effect of shear stress on expression of a recombinant protein by Chinese hamster ovary cells. Biotechnol Bioeng 2003;81:211–220.
25. Dardik A, et al. Shear stress-stimulated endothelial cells induce smooth muscle cell chemotaxis via platelet-derived growth factor-BB and interleukin-1 alpha. J Vasc Surg 2005;41:321–331.
26. Kher N, Marsh JD. Pathobiology of atherosclerosis—a brief review. Semin Thromb Hemostasis 2004;30:665–672.
27. Butcher JT, Penrod AM, Garcia AJ, Nerem RM. Unique morphology and focal adhesion development of valvular endothelial cells in static and fluid flow environments. Arterioscler Thromb Vasc Biol 2004;24:1429–1434.
28. Kladakis SM, Nerem RM. Endothelial cell monolayer formation: Effect of substrate and fluid shear stress. Endothelium 2004;11:29–44.
29. Imberti B, Seliktar D, Nerem RM, Remuzzi A. The response of endothelial cells to fluid shear stress using a co-culture model of the arterial wall. Endothelium-New York 2002;9: 11–23.
30. Song J, et al. A Computer-Controlled Microcirculatory Support System for Endothelial Cell Culture and Shearing. Anal Chem 2005, In press.
31. Krizanac-Bengez L, Mayberg MR, Janigro D. The cerebral vasculature as a therapeutic target for neurological disorders and the role of shear stress in vascular homeostatis and pathophysiology. Neurol Res 2004;26:846–853.
32. Richards M, et al. Bone regeneration and fracture healing—Experience with distraction osteogenesis model. Clin Orthop Related Res 1998; S191–S204.
33. Kim MB, Sarelius IH. Role of shear forces and adhesion molecule distribution on P-selectin-mediated leukocyte rolling in postcapillary venules. Amer J Physiol-Heart Circ Phy 2004;287:H2705–H2711.
34. Blackwell TS, Christman JW. The role of nuclear factor-kappa B in cytokine gene regulation. Amer J Respir Cell Molec Biol 1997;17:3–9.
35. Awolesi MA, Sessa WC, Sumpio BE. Cyclic Strain up-Regulates Nitric-Oxide Synthase in Cultured Bovine Aortic Endothelial-Cells. J Clin Invest 1995;96:1449–1454.
36. Suzuki N, et al. Up-regulation of integrin beta (3) expression by cyclic stretch in human umbilical endothelial cells. Biochem Biophys Res Commun 1997;239:372–376.
37. Booz GW, Baker KM. Molecular signaling mechanisms controlling growth and function of cardiac fibroblasts. Cardiovasc Res 1995;30:537–543.
38. Kolpakov V, et al. Effect of mechanical forces on growth and matrix protein-synthesis in the in-vitro pulmonary-artery—analysis of the role of individual cell-types. Circ Res 1995;77:823–831.
39. Desrosiers EA, Methot S, Yahia LH, Rivard CH. Response of ligamentous fibroblasts to mechanical stimulation. Ann Chir 1995;49:768–774.
40. Komuro I, et al. Mechanical loading stimulates cell hypertrophy and specific gene-expression in cultured rat cardiac myocytes—possible role of protein-kinase-C activation. J Biol Chem 1991;266:1265–1268.
41. Neidlingerwilke C, Wilke HJ, Claes L. Cyclic stretching of human osteoblasts affects proliferation and metabolism—a new experimental—method and its application. J Orthopaed Res 1994;12:70–78.
42. Vandenburgh HH, Karlisch P. Longitudinal growth of skeletal myotubes in vitro in a new horizontal mechanical cell stimulator. In Vitro Cell Develop Biol 1989;25:607–616.
43. Reusch P, et al. Mechanical strain increases smooth muscle and decreases nonmuscle myosin expression in rat vascular smooth muscle cells. Circ Res 1996;79:1046–1053.
44. Stauber WT, Miller GR, Grimmett JG, Knack KK. Adaptation of rat soleus muscles to 4-wk of intermittent strain. J Appl Physiol 1994;77:58–62.
45. Gudi SRP, Lee AA, Clark CB, Frangos JA. Equibiaxial strain and strain rate stimulate early activation of G proteins in cardiac fibroblasts. Amer J Physiol-Cell Physiol 1998; 43:C1424–C1428.
46. Vandenburgh HH, Hatfaludy S, Sohar I, Shansky J. Stretch-induced prostaglandins and protein-turnover in cultured skeletal-muscle. Amer J Physiol 1990;259:C232–C240.
47. Cowan DB, Lye SJ, Langille BL. Regulation of vascular connexin43 gene expression by mechanical loads. Circ Res 1998;82:786–793.
48. Davies PF, Flow-mediated endothelial mechanotransduction. Physiol Rev 1995;75:519–560.
49. Blawas AS, Reichert WM. Protein patterning. Biomaterials 1998;19:595–609.
50. Zhu XY, et al. Fabrication of reconfigurable protein matrices by cracking. Nat Mater 2005.
51. Zhu XY, Bersano-Begey TF, Takayama S. Nanomaterials for Cell Engineering. In: Nalwa HS. editor, Encyclopedia of Nanoscience and Nanotechnology. American Scientific Publishers; 2004. p 857—878.
52. Ratner BD, Bryant SJ. Biomaterials: Where we have been and where we are going. Annu Rev Biomed Eng 2004;6:41–75.
53. Whitesides GM, et al. Soft lithography in biology and biochemistry. Annu Rev Biomed Eng 2001;3:335–373.
54. Delamarche E, et al. Microfluidic networks for chemical patterning of substrate: Design and application to bioassays. J Am Chem Soc 1998;120:500–508.
55. Takayama S, et al. Patterning cells and their environments using multiple laminar fluid flows in capillary networks. Proc Natl Acad Sci U S A 1999;96:5545–5548.
56. Folch A, Toner M. Cellular micropatterns on biocompatible materials. Biotechnol Prog 1998;14:388–392.

57. Grigioni M, et al. Pulsatile flow and atherogenesis: Results from in vivo studies. Int J Artif Organs 2001;24:784–792.
58. Satcher RL, Bussolari SR, Gimbrone MA, Dewey CF. The distribution of fluid forces on model arterial endothelium using computational fluid-dynamics. J Biomech Eng 1992;114:309–316.
59. Evans DJR, Britland S, Wigmore PM. Differential response of fetal and neonatal myoblasts to topographical guidance cues in vitro. Dev Genes Evol 1999;209:438–442.
60. Seemann R, et al. Wetting morphologies at microstructured surfaces. Proc Natl Acad Sci USA 2005;102:1848–1852.
61. Ogbonna JC, Tanaka H. Night biomass loss and changes in biochemical composition of cells during light/dark cyclic culture of Chlorella pyrenoidosa. J Ferment Bioeng 1996;82: 558–564.
62. Macleod RM, Dunn F. Ultrasonic irradiation of enzyme solutions. J Acoust Soc Amer 1966;40:1202.
63. Ishimori Y, Karube I, Suzuki S. Acceleration of immobilized alpha-chymotrypsin activity with ultrasonic irradiation. J Mol Catal 1981;12:253–259.
64. Vulfson EN, Sarney DB, Law BA. Enhancement of subtilisin-catalyzed interesterification in organic-solvents by ultrasound irradiation. Enzyme Microb Technol 1991;13:123–126.
65. Wood BE, Aldrich HC, Ingram LO. Ultrasound stimulates ethanol production during the simultaneous saccharification and fermentation of mixed waste office paper. Biotechnol Prog 1997;13:232–237.
66. Ostrovidov S, Jiang JL, Sakai Y, Fujii T. Membrane-based PDMS microbioreactor for perfused 3D primary rat hepatocyte cultures. Biomed Microdevices 2004;6:279–287.
67. Geiger RV, Berk BC, Alexander RW, Nerem RM. Flow-induced calcium transients in single endothelial-cells—spatial and temporal analysis. Amer J Physiol 1992;262:C1411–C1417.
68. Schneegass I, Brautigam R, Kohler JM. Miniaturized flow-through PCR with different template types in a silicon chip thermocycler. Lab Chip 2001;1:42–49.
69. Shivashankar GV, Liu S, Libchaber A. Control of the expression of anchored genes using micron scale heater. Appl Phys Lett 2000;76:3638–3640.
70. Shen KY, Chen XF, Guo M, Cheng J. A microchip-based PCR device using flexible printed circuit technology. Sensors and Actuators B-Chem 2005;105:251–258.
71. Northrup MA, Ching MT, White RM, Watson RT. DNA amplification with a microfabricated reaction chamber. Proc IEEE Int Conf Solid-state Sensors Actuators 1993; 924–926.
72. Kopp MU, de Mello AJ, Manz A. Chemical amplification: Continuous-flow PCR on a chip. Science 1998;280:1046–1048.
73. Krishnan M, Ugaz VM, Burns MA. PCR in a Rayleigh-Benard convection cell. Science 2002;298:793–793.
74. Elwenspoek M, Lammerink TSJ, Miyake R, Fluitman JHJ. Towards integrated microliquid handling systems. J Micromechanic Microengineer 1994;4:227–245.
75. Lao AIK, Lee TMH, Hsing IM, Ip NY. Precise temperature control of microfluidic chamber for gas and liquid phase reactions. Sensors and Actuators A-Phys 2000;84:11–17.
76. Altman GH, et al. Advanced bioreactor with controlled application of multi-dimensional strain for tissue engineering. J Biomech Eng 2002;124:742–749.
77. Thorsen T, Maerkl SJ, Quake SR. Microfluidic large-scale integration. Science 2002;298:580–584.
78. Grover WH, et al. Monolithic membrane valves and diaphragm pumps for practical large-scale integration into glass microfluidic devices. Sensors and Actuators B-Chem 2003;89: 315–323.
79. Vandepol FCM, Wonnink DGJ, Elwenspoek M, Fluitman JHJ. A thermo-pneumatic actuation principle for a miniature pump and other micromechanical devices. Sensors and Actuators 1989;17:139–143.
80. Smits GJ. A piezoelectric micropump with three valves working peristalitically. Sensors and Actuators 1990;15:203–206.
81. Bosch D, et al. A silicon microvalve with combined electromagnetic/electrostatic actuation. Sensors and Actuators A-Phys 1993;37-8:684–692.
82. Liu RH, Yu Q, Beebe DJ. Fabrication and characterization of hydrogel-based microvalves. J Microelectromech Syst 2002; 11:45–53.
83. Ehrick JD, et al. Genetically engineered protein in hydrogels tailors stimuli-responsive characteristics. Nat Mater 2005;4: 298–302.
84. Gu W, et al. Computerized microfluidic cell culture using elastomeric channels and Braille displays. Proc Natl Acad Sci U S A 2004;101:15861–15866.
85. Unger MA, et al. Monolithic microfabricated valves and pumps by multilayer soft lithography. Science 2000;288: 113–116.
86. Lemoff AV, Lee AP. An AC magnetohydrodynamic micropump. Sens Actuators B-Chem 2000;63:178–185.
87. Zhu XY, et al. Arrays of horizontally-oriented mini-reservoirs generate steady microfluidic flows for continuous perfusion cell culture and gradient generation. Analyst 2004;129:1026–1031.
88. Takayama S, et al. Selective chemical treatment of cellular microdomains using multiple laminar streams. Chem Biol 2003;10:123–130.
89. Dertinger SKW, Chiu DT, Jeon NL, Whitesides GM. Generation of gradients having complex shapes using microfluidic networks. Anal Chem 2001;73:1240–1246.
90. Jeon NL, et al. Neutrophil chemotaxis in linear and complex gradients of interleukin-8 formed in a microfabricated device. Nat Biotechnol 2002;20:826–830.
91. Hatch A, et al. A rapid diffusion immunoassay in a T-sensor. Nat Biotechnol 2001;19:461–465.
92. Sawano A, Takayama S, Matsuda M, Miyawaki A. Lateral propagation of EGF signaling after local stimulation is dependent on receptor density. Dev Cell 2002;3:245–257.
93. Bessoth FG, deMello AJ, Manz A. Microstructure for efficient continuous flow mixing. Anal Commun 1999;36:213–215.
94. Hong JW, et al. A nanoliter-scale nucleic acid processor with parallel architecture. Nat Biotechnol 2004;22:435–439.
95. Liu RH, et al. Passive mixing in a three-dimensional serpentine microchannel. J Microelectromech Syst 2000;9:190–197.
96. Song H, Tice JD, Ismagilov RF. A microfluidic system for controlling reaction networks in time. Angew Chem Int Ed Engl 2003;42:768–772.
97. Burke BJ, Regnier FE. Stopped-flow enzyme assays on a chip using a microfabricated mixer. Anal Chem 2003;75:1786–1791.
98. Hessel V, Lowe H, Schonfeld F. Micromixers—a review on passive and active mixing principles. Chem Eng Sci 2005;60:2479–2501.
99. Nguyen NT, Wu ZG. Micromixers—a review. J Micromechanic Microengineer 2005;15:R1–R16.
100. Frangos JA, McIntire LV, Eskin SG. Shear-Stress Induced stimulation of mammalian-cell metabolism. Biotechnol Bioeng 1988;32:1053–1060.
101. Blackman BR, Barbee KA, Thibault LE. In vitro cell shearing device to investigate the dynamic response of cells in a controlled hydrodynamic environment. Ann Biomed Eng 2000;28:363–372.
102. Frame MDS, Chapman GB, Makino Y, Sarelius IH. Shear stress gradient over endothelial cells in a curved microchannel system. Biorheology 1998;35:245–261.

103. Frame MD, Sarelius IH. Flow-induced cytoskeletal changes in endothelial cells growing on curved surfaces. Microcirculation 2000;7:419–427.
104. Chien S, Li S, Shyy JYJ. Effects of mechanical forces on signal transduction and gene expression in endothelial cells. Hypertension 1998;31:162–169.
105. Clark CB, Burkholder TJ, Frangos JA. Uniaxial strain system to investigate strain rate regulation in vitro. Rev Sci Instr 2001;72:2415–2422.
106. Buck RC. Reorientation response of cells to repeated stretch and recoil of the substratum. Exp Cell Res 1980;127:470–474.
107. Mourgeon E, et al. Mechanical strain-induced posttranscriptional regulation of fibronectin production in fetal lung cells. Amer J Physiol-Lung Cell M Ph 1999;277:L142–L149.
108. Kato T, et al. Up-regulation of COX2 expression by uni-axial cyclic stretch in human lung fibroblast cells. Biochem Biophys Res Commun 1998;244:615–619.
109. Schilling EA, Kamholz AE, Yager P. Cell lysis and protein extraction in a microfluidic device with detection by a fluorogenic enzyme assay. Anal Chem 2002;74:1798–1804.
110. Soughayer JS, et al. Characterization of cellular optoporation with distance. Anal Chem 2000;72:1342–1347.
111. Lee SW, Tai YC. A micro cell lysis device. Sens Actuators A-Phys 1999;73:74–79.
112. Heo J, Thomas KJ, Seong GH, Crooks RM. A microfluidic bioreactor based on hydrogel-entrapped E. coli: Cell viability, lysis, and intracellular enzyme reactions. Anal Chem 2003;75:22–26.
113. Waters LC, et al. Microchip device for cell lysis, multiplex PCR amplification, and electrophoretic sizing. Anal Chem 1998;70:158–162.
114. Huang Y, et al. Electric manipulation of bioparticles and macromolecules on microfabricated electrodes. Anal Chem 2001;73:1549–1559.
115. Leitz G, Weber G, Seeger S, Greulich KO. The laser microbeam trap as an optical tool for living cells. Physiol Chem Phys Med Nmr 1994;26:69–88.
116. Hogman CF. Preparation and preservation of red cells. Vox Sang 1998;74:177–187.
117. Bauer J. Advances in cell separation: Recent developments in counterflow centrifugal elutriation and continuous flow cell separation. J Chromatogr B 1999;722:55–69.
118. Cheng J, Kricka LJ, Sheldon EL, Wilding P. Sample preparation in microstructured devices, microsystem technology in chemistry and life science. Top Curr Chem 1998; 215–231.
119. Fu AY, et al. A microfabricated fluorescence-activated cell sorter. Nat Biotechnol 1999;17:1109–1111.
120. Rathman M, et al. The development of a FACS-based strategy for the isolation of Shigella flexneri mutants that are deficient in intercellular spread. Mol Microbiol 2000;35:974–990.
121. Handgretinger R, et al. Isolation and transplantation of autologous peripheral CD34(+) progenitor cells highly purified by magnetic-activated cell sorting. Bone Marrow Transplant 1998;21:987–993.
122. Volkmuth WD, Austin RH. DNA electrophoresis in microlithographic arrays. Nature(London) 1992;358:600–602.
123. Li PCH, Harrison DJ. Transport, manipulation, and reaction of biological cells on-chip using electrokinetic effects. Anal Chem 1997;69:1564–1568.
124. Wang XB, et al. Cell separation by dielectrophoretic field-flow-fractionation. Anal Chem 2000;72:832–839.
125. Cho BS, et al. Passively driven integrated microfluidic system for separation of motile sperm. Anal Chem 2003;75: 1671–1675.
126. Horsman KM, et al. Separation of sperm and epithelial cells in a microfabricated device: Potential application to forensic analysis of sexual assault evidence. Anal Chem 2005;77: 742–749.
127. Chou CF, et al. Electrodeless dielectrophoresis of single- and double-stranded DNA. Biophys J 2002;83:2170–2179.
128. Shevkoplyas SS, Yoshida T, Munn LL, Bitensky MW. Biomimetic autoseparation of leukocytes from whole blood in a microfluidic device. Anal Chem 2005;77:933–937.
129. Petersson F, et al. Continuous separation of lipid particles from erythrocytes by means of laminar flow and acoustic standing wave forces. Lab Chip 2005;5:20–22.
130. Minc N, Viovy JL. Microfluidics and biological applications: the stakes and trends. C R Phys 2004;5:565–575.
131. Huh D, et al. Microfluidics for flow cytometric analysis of cells and particles. Physiol Meas 2005;26:R1–R26.
132. Zhao B, Viernes NOL, Moore JS, Beebe DJ. Control and applications of immiscible liquids in microchannels. J Am Chem Soc 2002;124:5284–5285.
133. Hisamoto H, et al. Chemicofunctional membrane for integrated chemical processes on a microchip. Anal Chem 2003; 75:350–354.
134. Kawabata T, Washizu M. Dielectrophoretic detection of molecular bindings. IEEE Trans Ind Appl 2001;37:1625–1633.
135. Wu XZ, Suzuki M, Sawada T, Kitamori T. Chemiluminescence on a microchip. Anal Sci 2000;16:321–323.
136. Roda A, et al. Biotechnological applications of bioluminescence and chemiluminescence. Trends Biotech 2004;22:295–303.
137. Johnson ME, Landers JP. Fundamentals and practice for ultrasensitive laser-induced fluorescence detection in microanalytical systems. Electrophoresis 2004;25:3513–3527.
138. Mogensen KB, Klank H, Kutter JP. Recent developments in detection for microfluidic systems. Electrophoresis 2004;25: 3498–3512.
139. Xu H, et al. A real-time ratiometric method for the determination of molecular oxygen inside living cells using sol-gel-based spherical optical nanosensors with applications to rat C6 glioma. Anal Chem 2001;73:4124–4133.
140. Wolfbeis OS. Fiber-optic chemical sensors and biosensors. Anal Chem 2002;74:2663–2677.
141. Chang-Yen DA, Gale BK. An integrated optical oxygen sensor fabricated using rapid-prototyping techniques. Lab Chip 2003;3:297–301.
142. Vandaveer WR, et al. Recent developments in electrochemical detection for microchip capillary electrophoresis. Electrophoresis 2004;25:3528–3549.
143. Bakker E, Telting-Diaz M. Electrochemical sensors. Anal Chem 2002;74:2781–2800.
144. Walther I, et al. Development of a miniature bioreactor for continuous-culture in a space laboratory. J Biotechnol 1994;38:21–32.
145. Yoon HJ, et al. Solid-state ion sensors with a liquid junction-free polymer membrane-based reference electrode for blood analysis. Sens Actuators B-Chem 2000; 64:8–14.
146. Johannessen EA, et al. Micromachined nanocalorimetric sensor for ultra-low-volume cell-based assays. Anal Chem 2002;74:2190–2197.
147. Li PCH, Wang WJ, Parameswaran M. An acoustic wave sensor incorporated with a microfluidic chip for analyzing muscle cell contraction. Analyst 2003;128:225–231.
148. Kurita R, et al. Differential measurement with a microfluidic device for the highly selective continuous measurement of histamine released from rat basophilic leukemia cells (RBL-2H3). Lab Chip 2002;2:34–38.

149. Matsubara Y, et al. Application of on-chip cell cultures for the detection of allergic response. Biosens Bioelectron 2004;19: 741–747.
150. Daunert S, et al. Genetically engineered whale-cell sensing systems: Coupling biological recognition with reporter genes. Chem Rev 2000;100:2705–2738.
151. Wu XZ, et al. Immunofluorescent labeling of cancer marker her2 and other cellular targets with semiconductor quantum dots. Nat Biotechnol 2003;21:41–46.
152. Parak WJ, Pellegrino T, Plank C. Labelling of cells with quantum dots. Nanotechnology 2005;16:R9–R25.
153. Cui Y, Wei QQ, Park HK, Lieber CM. Nanowire nanosensors for highly sensitive and selective detection of biological and chemical species. Science 2001;293:1289–1292.
154. Chou SY, Krauss PR, Renstrom PJ. Imprint lithography with 25-nanometer resolution. Science 1996;272:85–87.
155. Raiteri R, Grattarola M, Butt HJ, Skladal P. Micromechanical cantilever-based biosensors. Sens Actuators B-Chem 2001;79:115–126.
156. Moriguchi H, et al. An agar-microchamber cell-cultivation system:flexible change of microchamber shapes during cultivation by photo-thermal etching. Lab Chip 2002;2:125–130.
157. Burns MA, et al. An integrated nanoliter DNA analysis device. Science 1998;282:484–487.
158. Hansen CL, Skordalakes E, Berger JM, Quake SR. A robust and scalable microfluidic metering method that allows protein crystal growth by free interface diffusion. Proc Natl Acad Sci USA 2002;99:16531–16536.
159. Chou HP, Spence C, Scherer A, Quake S. A microfabricated device for sizing and sorting DNA molecules. Proc Natl Acad Sci USA 1999;96:11–13.
160. DeBusschere BD, Kovacs GTA. Portable cell-based biosensor system using integrated CMOS cell-cartridges. Biosens Bioelectron 2001;16:543–556.
161. Erickson D, Li DQ. Integrated microfluidic devices. Anal Chim Acta 2004;507:11–26.
162. Shito M, et al. In vitro and in vivo evaluation of albumin synthesis rate of porcine hepatocytes in a flat-plate bioreactor. Artif Organs 2001;25:571–578.
163. Tilles AW, et al. Effects of oxygenation and flow on the viability and function of rat hepatocytes cocultured in a microchannel flat-plate bioreactor. Biotechnol Bioeng 2001;73:379–389.
164. Nyberg SL, et al. Primary hepatocytes outperform Hep G2 cells as the source of biotransformation functions in a bioartificial liver. Ann Surg 1994;220:59–67.
165. Powers MJ, et al. Functional behavior of primary rat liver cells in a three-dimensional perfused microarray bioreactor. Tissue Eng 2002;8:499–513.
166. Bader A, et al. Development of a small-scale bioreactor for drug metabolism studies maintaining hepatospecific functions. Xenobiotica 1998;28:815–825.
167. Allen JW, Bhatia SN. Improving the next generation of bioartificial liver devices. Semin Cell Dev Biol 2002; 13:447–454.
168. Ghanem A, Shuler ML. Combining cell culture analogue reactor designs and PBPK models to probe mechanisms of naphthalene toxicity. Biotechnol Prog 2000;16:334–345.
169. Ghanem A, Shuler ML. Characterization of a perfusion reactor utilizing mammalian cells on microcarrier beads. Biotechnol Prog 2000;16:471–479.
170. Park TH, Shuler ML. Integration of cell culture and microfabrication technology. Biotechnol Prog 2003;19:243–253.
171. Viravaidya K, Shuler ML. Incorporation of 3T3-L1 cells to mimic bioaccumulation in a microscale cell culture analog device for toxicity studies. Biotechnol Prog 2004;20:590–597.
172. Kruip TAM, Bevers MM, Kemp B. Environment of oocyte and embryo determines health of IVP offspring. Theriogenology 2000;53:611–618.
173. Bavister BD. Interactions between embryos and the culture milieu. Theriogenology 2000;53:619–626.
174. Fukui Y, Lee ES, Araki N. Effect of medium renewal during culture in two different culture systems on development to blastocysts from in vitro produced early bovine embryos. J Anim Sci 1996;74:2752–2758.
175. Beebe D, et al. Microfluidic technology for assisted reproduction. Theriogenology 2002;57:125–135.
176. Glasgow IK, et al. Handling individual mammalian embryos using microfluidics, IEEE Trans Biomed Eng 2001;48:570–578.
177. Zeringue HC, Wheeler MB, Beebe DJ. Zona pellucida removal of mammalian embryos in a microfluidic systems. Micro Total Analysis Syst 2000; 214–217.
178. Zeringue HC, Beebe DJ, Wheeler MB. Removal of cumulus from mammalian zygotes using microfluidic techniques. Biomed Microdevices 2001;3:219–224.
179. Hickman DL, Beebe DJ, Rodriguez-Zas SL, Wheeler MB. Comparison of static and dynamic medium environments for culturing of pre-implantation mouse embryos. Comparative Med 2002;52:122–126.
180. Suh RS, et al. Rethinking gamete/embryo isolation and culture with microfluidics. Hum Reprod Update 2003;9:451–461.
181. Korbutt GS, et al. Improved survival of microencapsulated islets during in vitro culture and enhanced metabolic function following transplantation. Diabetologia 2004;47:1810–1818.
182. Vallbacka JJ, Nobrega JN, Sefton MV. Tissue engineering as a platform for controlled release of therapeutic agents: implantation of microencapsulated dopamine producing cells in the brains of rats. J Control Release 2001;72:93–100.
183. Takenaga M, et al. A single treatment with microcapsules containing a CXCR4 antagonist suppresses pulmonary metastasis of murine melanoma. Biochem Biophys Res Commun 2004;320:226–232.
184. Yu BL, Chang TMS. Effects of long-term oral administration of polymeric microcapsules containing tyrosinase on maintaining decreased systemic tyrosine levels in rats. J Pharm Sci 2004;93:831–837.
185. AlHendy A, Hortelano G, Tannenbaum GS, Chang PL. Growth retardation—an unexpected outcome from growth hormone gene therapy in normal mice with microencapsulated myoblasts. Hum Gene Ther 1996;7:61–70.
186. Zilberman Y, et al. Polymer-encapsulated engineered adult mesenchymal stem cells secrete exogenously regulated rhBMP-2, and induce osteogenic and angiogenic tissue formation. Polym Adv Technol 2002;13:863–870.
187. Cirone P, Bourgeois JM, Austin RC, Chang PL. A novel approach to tumor suppression with microencapsulated recombinant cells. Hum Gene Ther 2002;13:1157–1166.
188. Hortelano G, Wang L, Xu N, Ofosu FA. Sustained and therapeutic delivery of factor IX in nude haemophilia B mice by encapsulated C2C12 myoblasts: Concurrent tumourigenesis. Haemophilia 2001;7:207–214.
189. Chen JP, et al. Microencapsulation of islets in PEG-amine modified alginate-poly(L-lysine)-alginate microcapsules for constructing bioartificial pancreas. J Ferment Bioeng 1998;86:185–190.
190. Xue YL, et al. Pain relief by xenograft of subarachnoid microencapsulated bovine chromaffin cells in cancer patients. Prog Nat Sci 2000;10:919–924.

191. Cole DR, et al. Transplantation of microcapsules (a potential bioartificial organ)—biocompatibility and host-reaction. J Mater Sci-Mater Med 1993;4:437–442.
192. Lou WH, Qin XY, Wu ZG. Preliminary research on biocompatibility of alginate-chitosan-polyethyleneglycol microcapsules. Minerva Biotecnol 2000;12:235–240.
193. Van Raamsdonk JM, Cornelius RM, Brash JL, Chang PL. Deterioration of polyamino acid-coated alginate microcapsules in vivo. J Biomater Sci-Polym Ed 2002;13:863–884.
194. Sakai S, Ono T, Ijima H, Kawakami K. Behavior of enclosed sol- and gel-alginates in vivo. Biochem Eng J 2004;22:19–24.
195. Koch S, et al. Alginate encapsulation of genetically engineered mammalian cells:comparison of production devices, methods and microcapsule characteristics. J Microencapsul 2003;20:303–316.
196. Arica B, et al. Carbidopa/levodopa-loaded biodegradable microspheres:in vivo evaluation on experimental Parkinsonism in rats. J Control Release 2005;102:689–697.
197. Leblond FA, et al. Studies on smaller (similar to 315 (μM) microcapsules: IV. Feasibility and safety of intrahepatic implantations of small alginate poly-L-lysine microcapsules. Cell Transplant 1999;8:327–337.
198. Chang TMS, Prakash S. Therapeutic uses of microencapsulated genetically engineered cells. Mol Med Today 1998;4: 221–227.
199. Jeon NL, et al. Whitesides GM. Generation of solution and surface gradients using microfluidic systems. Langmuir 2000;16:8311–8316.
200. Sprague EA, Steinbach BL, Nerem RM, Schwartz CJ. Influence of a laminar steady-state fluid-imposed wall shear-stress on the binding, internalization, and degradation of low-density lipoproteins by cultured arterial endothelium. Circulation 1987;76:648–656.
201. Diamond SL, Eskin SG, McIntire LV. Fluid-flow stimulates tissue plasminogen-activator secretion by cultured human-endothelial cells. Science 1989;243:1483–1485.
202. Levesque MJ, Sprague EA, Schwartz CJ, Nerem RM. The influence of shear-stress on cultured vascular endothelial-cells—the stress response of an anchorage-dependent mammalian-cell. Biotechnol Prog 1989;5:1–8.
203. Gomes N, et al. Shear stress modulates tumour cell adhesion to the endothelium. Biorheology 2003;40:41–45.
204. Davies PF, Tripathi SC. Mechanical-stress mechanisms and the cell—an endothelial paradigm. Circ Res 1993;72:239–245.
205. Ikeda M, et al. Extracellular signal-regulated kinases 1 and 2 activation in endothelial cells exposed to cyclic strain. Am J Physiol-Heart Circul Physiol 1999;276: H614–H622.
206. Smith PG, Roy C, Zhang YN, Chauduri S. Mechanical stress increases RhoA activation in airway smooth muscle cells. Am J Respir Cell Mol Bio 2003;28:436–442.
207. Lee T, Kim SJ, Sumpio BE. Role of PP2A in the regulation of p38-MAPK activation in bovine aortic endothelial cells exposed to cyclic strain. J Cell Physiol 2003;194:349–355.
208. Han O, Takei T, Basson M, Sumpio BE. Translocation of PKC isoforms in bovine aortic smooth muscle cells exposed to strain. J Cell Biochem 2001;80:367–372.
209. Shelby JP, et al. A microfluidic model for single-cell capillary obstruction by Plasmodium falciparum infected erythrocytes. Proc Natl Acad Sci USA 2003;100:14618–14622.

See also MICROARRAYS; MICROFLUIDICS; NANOPARTICLES; TISSUE ENGINEERING.

MICRODIALYSIS SAMPLING

JULIE A. STENKEN
Rensselaer Polytechnic Institute
Troy, New York

MICRODIALYSIS SAMPLING: NON-SPECIALIST VIEW

Microdialysis sampling devices are minimally invasive miniature dialyzers that can be implanted into a distinct tissue region to obtain a chemical snapshot over an integrated time period. In combination with appropriate chemical detection methods for the targeted substances, a microdialysis sampling device may be considered to be a universal biosensor. Obtaining chemical information from different tissues can often lead to either a greater understanding of the underlying chemistry involved with the physiological function of the organ or the origin of a particular disease process. A simplified view of the microdialysis sampling device is shown in Fig. 1. The central part of the microdialysis sampling device is a single semipermeable hollow fiber membrane with dimensions that range between 200 and 500 μm for its external diameter and 1 and 30 mm in length. A perfusion solution is passed through the device at microliter per minute flow rates. Compounds diffuse from the tissue space into the dialysis probe and are carried to an outlet to undergo chemical analysis. Originally microdialysis sampling was developed to obtain real-time chemical information from rodent brain and was termed intracranial dialysis. Microdialysis sampling has now been applied for chemical collection from nearly every single organ. In addition to neurotransmitter collection, the device has been used for endocrinology, immunology, metabolism, and pharmacokinetic applications as shown in Table 1. The biomedical literature cites

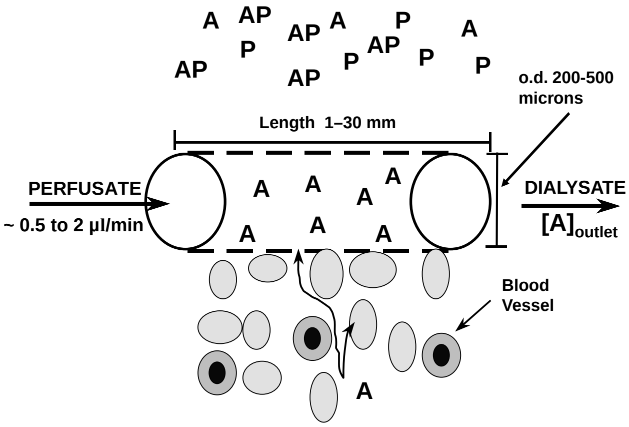


Figure 1. Microdialysis sampling process. A perfusion fluid that closely matches the ionic strength and composition of the fluid external to the microdialysis membrane is passed through at flow rates between 0.5 and 2.0 $\mu L{\cdot}min^{-1}$. Analytes, A, that are not protein bound, AP, diffuse through the extracellular fluid space (wavy lines) and can pass through the pores of the semipermeable membrane are collected into the device. The analyte outlet concentration $[A]_{outlet}$ can then be quantified using a suitable detection method.

Table 1. Typical Microdialysis Sampling Uses

Collection of endogenous analytes from brain including neurotransmitters (dopamine, norepinephrine, serotonin, glutamate, GABA[a], glucose, peptides and proteins (cytokines).
Collection of endogenous small hydrophilic analytes from other tissues (e.g., glucose for diabetics).
Collection of peptides and proteins from perphiral tissues, for example, glutathione, neuropeptides, and different cytokines and growth factors (e.g., VEGF).
Collection of xenobiotic analytes for pharmacokinetic or pharmacodynamic studies from numerous tissue sites (brain, dermis, muscle, and tumors) in both animals and humans.
Localized delivery of analytes followed by concomitant recovery of endogenous analytes or metabolized products (e.g., spin traps, metabolites).

[a] GABA = r-aminobutyric acid

thousands of research articles that have used microdialysis sampling devices to study many different basic and clinical research problems with perhaps 90% or greater of these applications focused in neuroscience. As more life scientists realize that microdialysis sampling devices exist, this device will be used more often to solve many additional clinical problems outside of the neurosciences.

INTRODUCTION

Mammalian organs are highly complex systems that achieve their functions through a multifaceted chemical communication network. For laboratory studies focused on understanding these chemical networks, the organ is often broken down into its component parts (cells, subcellular components, extracellular matrix components, etc.) to create more controlled conditions. These types of studies have allowed for a great understanding of the individual component parts, but do not address how the chemical communication may occur within the intact organ. Common diagnostic collection methods, such as blood or urine sampling, are too far removed from organs to be able to provide desired information about localized organ biochemistry. Gaining chemical information from an organ system prior to the creation of microdialysis sampling devices required either organ dissection or noninvasive analysis methods. Removing organs to access chemical content is fraught with many concerns including the preparation of the sample that involves in most cases sacrifice of the animal (or a biopsy for humans) as well as concerns about chemical stability and loss during the sample preparation process. For nearly all organs, the changes in localized tissue chemistry caused during sacrifice may severely alter some chemical communication systems. Finally, organ dissection does not allow for temporal studies of targeted analytes.

An increasing number of medical devices and instruments are becoming available to noninvasively measure *in vivo* chemical composition at spatially defined sites. Noninvasive measurements typically use spectroscopic instruments that are highly analyte specific. In particular, the most well-known medical devices for achieving these tasks are positron emission tomography (PET) and magnetic resonance imaging (MRI). Positron emission tomography scanning requires production of special isotopes (^{11}C, ^{13}N, ^{15}O, ^{18}F) with limited half-lives. Similarly, MRI uses ^{1}H to create an image and is a useful imaging technique because the concentration of hydrogen nuclei in water and fat is roughly 100 *M* (mol·/L^{-1}). Magnetic resonance spectroscopy (MRS) can be used to noninvasively detect other isotopes (^{1}H, ^{13}C, ^{15}N, ^{19}F, ^{31}P), but requires very high concentrations of these nuclei for detection. Fluorescence imaging has also been used for noninvasive measurements in conjunction with near-infrared (IR) tags or with enzyme substrates that become fluorescent when cleaved (1–3). While these spectroscopic techniques hold significant promise for some areas of clinical medicine (cardiovascular and oncology), they are quite limited with respect to the range of analytes that can be detected and thus potential applications. While not noninvasive, minimally invasive microdialysis sampling devices coupled with appropriate dialysate analytical detection methods allow for measurements of localized tissue chemistry with significantly greater analyte flexibility and spatial resolution.

Microdialysis sampling originated as an alternative to push–pull perfusion devices for use in mammalian brain. Push–pull perfusion devices were used to collect fluid relevant to synaptic transmission (4). During a push–pull perfusion, a solution that closely matches the ionic chemical composition of the extracellular fluid (ECF) is loaded into a syringe and this is gently infused into the implantation site. This perfusion fluid mixes with the existing ECF and then is pulled back into the device. The analysis of push–pull samples became a useful tool for tracking neurotransmitter activities coupled with allowing for a remarkable understanding of the underlying chemical events associated with a wide variety of behavioral and physiological stimuli. Insertion of push–pull cannula could potentially cause tissue damage or lesions, which raised many concerns among researchers since this type of damage could limit the usefulness of the neurochemical data collected. In other words, researchers were often concerned that sampling was really taking place in a "lake" of fluid that may not contain the representative chemical components of the extracellular fluid space.

The push–pull sampling method was principally applied to neurochemical sampling. For other tissues, other extracellular fluid sampling methods have been described. These other methods include open-flow microperfusion and wick methods. Both of these methods have been used in muscle, as well as dermal sampling. Additionally, blister methods are common for obtaining interstitial fluid from the skin. Open-flow microperfusion is similar to push–pull perfusion. A cannula device contains an inner tube and an outer tube with open pores (typically $\sim$500 μm) is placed over this inner tube. A peristaltic pump then simultaneously delivers and withdraws fluid through the device (5). Open-flow microperfusion has primarily been used for sampling the extracellular fluid in muscle and skin. Alternatives to perfusion methods are wick methods, which use nylon wicks to sample the extracellular fluid space in animals and humans (6–8). To obtain multiple samples over a specified time period would require insertion and removal of individual wick devices. This repeated insertion and removal might cause additional trauma to the sampling site.

To overcome the tissue damage concerns associated with push–pull perfusion for neuroscience applications, the use of semipermeable dialysis membranes at the tips of the push–pull cannula were used (9,10). These original dialysis bags eventually led to flow-through microdialysis probes introduced by Pycock and Ungerstedt in 1974 (11). The advantage of microdialysis sampling over the push–pull cannula is that fluid is not pushed into sensitive brain tissue. Unlike a push–pull perfusion, microdialysis sampling is a continuous process that provides a sample that excludes many of the components from the ECF. This exclusion process serves to provide a relatively clean sample for chemical analysis. Today, this technique has been widely used by life scientists to attain site-specific access to numerous tissue sites to study and solve many problems, which include, but are not limited to the following applications: (*1*) To elucidate the role of different neurotransmitters and neuropeptides in specifically defined brain regions; (*2*) To collect glucose continuously over many days to give a better chemical picture to diabetes specialists to determine the efficacy of an insulin regimen for individual patients; (*3*) To collect energy metabolites (glucose, lactate, pyruvate) as well as administered drugs (morphine) to understand diseased energy metabolism and blood-brain barrier transport from head trauma patients; (*4*) To determine if an efficacious drug concentration is reaching a specific infection site or diseased tissue sites for antineoplastic therapy or antimicrobial therapy; (*5*) To determine pharmacokinetic parameters in single animals; (*6*) To determine metabolite formation and accumulation in various tissues after a drug dose in a single animal over time; (*7*) To determine the extent of drug blood–brain barrier permeation of new drugs in animal models; (*8*) To collect various endogenous peptides and proteins (e.g., cytokines and growth factors) from different peripheral sites in animals and humans (*12*).

To newcomers to this device, the principles of microdialysis sampling operation at first seem deceptively simple. At the most basic level, the microdialysis sampling device may be considered to behave as an artificially implanted blood vessel that allows free analyte diffusion into the inner fiber lumen. However, as will be discussed in this article, numerous considerations that involve both an understanding of the localized biology and physiology, as well as the underlying mass transport processes are essential to microdialysis sampling data interpretation.

MICRODIALYSIS SAMPLING PRINCIPLES OF OPERATION

In principle, as long as the microdialysis probe can be implanted, it can be used for sampling localized tissue biochemistry. Microdialysis sampling requires only a few pieces of equipment. This equipment is not cost-prohibitive, which is the reason that many different researchers can perform microdialysis sampling experiments in their own laboratories. For most experiments, the necessary equipment includes a perfusion pump to deliver the perfusion fluid and a microdialysis probe. In addition to the pump and probe, for animal studies, a device to hold either an anesthetized animal (e.g., stereotaxic unit for neuroscience procedures) or a bowl with appropriate swivels to prevent tangled tubing for freely moving animals may be necessary.

Microdialysis Sampling Instrumentation Components

The basic components needed to perform microdialysis sampling experiments in an awake-freely moving animals has been described (13). The components required for this type of experiment includes the microperfusion pump, an inlet and outlet fluid swivel that prevents the fluid lines from becoming tangled, and a bowl system to allow the animal to freely move. Additional components can include refrigerated fraction collectors to allow collection and storage of sensitive samples. Microperfusion pumps used to deliver the perfusion fluid through the microdialysis probe are capable of delivering volumetric flow rates between 0.1 and 20 $\mu L{\cdot}min^{-1}$. Flow rates between 0.5 and 2.0 $\mu L{\cdot}min^{-1}$ are commonly used during most microdialysis sampling experiments.

Microdialysis sampling perfusion fluids are chosen to closely match the ionic strength and composition of the external tissue extracellular fluid surrounding the microdialysis probe. Perfusion fluids passed through microdialysis sampling probes are a form Ringer's solution for which there are numerous published chemical compositions (14,15). Typical Ringer's solutions contain ~150 m*M* NaCl, 4 m*M* KCl, and 2.4 m*M* $CaCl_2$ and can also be supplemented with glucose and other ionic salts ($MgCl_2$). These solutions are used both to maintain fluid balance, as well as ion balance across the dialysis membrane. Maintaining fluid balance across the dialysis membrane is important so that large osmotic pressures are not created. Significant osmotic pressure differences will cause fluid to be gained or lost during microdialysis sampling (16). Fluid loss is often undesirable for analytical as well as biological reasons. From an analytical perspective, oftentimes the analysis requires a set volume. For example, a liquid chromatographic analysis may require 10 μL of sample and a standard enzyme-linked immunosorbent assay (ELISA) may require 100 μL of sample. From a biological perspective, fluid loss can be undesirable in some tissues that are particularly sensitive, such as the brain. Furthermore, brain tissue is also highly sensitive to ionic concentration alterations since such changes can alter neurotransmitter release (17). By maintaining an osmotic balance, the fundamental mass transport mechanism for moving an analyte from the extracellular fluid space (ECF) to the dialysate lumen is principally diffusion.

Probe Geometry

Microdialysis sampling is typically considered to be synonymous with the term intracranial dialysis sampling because of its neuroscience origins. The first intracranial dialysis device was a linear design that traversed longitudinally through different brain regions. A variety of different probe designs have been described in many different review articles (18–20). Microdialysis probe design has evolved to allow use of the device for biomedical applications beyond neuroscience. Linear geometry microdialysis sampling devices for neuroscience were not as useful as the push–pull cannula that could be inserted into known brain regions (e.g., striatum, hippocampus, *substantia nigra*) based on known stereotaxic coordinates that in some cases are < 1 mm wide in rat brain. To overcome this challenge for neurochemistry studies, more rigid cannula designs were created that can be inserted into specific brain regions and are now commercially available from a variety of sources (21).

As microdialysis sampling devices became a standard tool used by neuroscientists, researchers in other fields began to realize its great *in vivo* analysis potential. Principally, the use of the probes for collection of endogenous or xenobiotic components in blood and peripheral tissues became of interest (22). In these tissues, a rigid stainless steel cannula causes tissue damage and may make awake and freely moving experiments with animals quite difficult. Cannula designs using Teflon or fused silica are commonly used for to make flexible probes for either sampling in soft peripheral tissues (e.g., skin or liver) or for blood sampling. Linear probe designs have also been reintroduced after originally being applied to brain studies and are now used for insertion into soft peripheral tissues. An additional advantage of these flexible designs is they also allow for studies in awake and freely moving animals in peripheral tissues. Recent research interests in transgenic mice have forced the creation of smaller microdialysis sampling devices (23).

Probe Materials

Semipermeable hollow fiber membranes used for microdialysis sampling are the same as those used for kidney dialysis. Different polymeric semipermeable membranes have been used in microdialysis sampling probes and are listed in Table 2. Typical materials include cellulose-based membranes (cuprophan or cellulose acetate), polycarbonate/polyether blends, polyacrylonitrile, and polyethersulfone. These membranes span a wide range of molecular weight cutoffs (MWCO) from 5000 to 100,000 Da. Choice of the membrane to be used during microdialysis sampling requires both analyte molecular weight information, as well as where the probe will be implanted as some tissue regions (particularly in the brain) are too narrow for > 500 μm external diameter membranes.

Table 2. Commercially Available Microdialysis Membrane Dimensions[a]

	PC	PES	PAN	CUP
Outer radius, μm	250	250	170	120
Inner radius, μm	200	205	120	95
Wall thickness, μm	50	45	50	25
Molecular weight cutoff	20,000	100,000	29,000	6,000
Outer surface area, mm^2	6.28	6.28	4.27	3.01

[a] The data provided here is that given by the manufacturers of the microdialysis probes. It is not known if the radii are for dry or wet membranes. The abbreviations are as follows: PC = polyether/polycarbonate, PES = polyethersulfone, PAN = polyacrylonitrile (or AN-69), CUP = cuprophan. CMA Microdialysis, Inc sells PC, PES, and CUP membranes. Bioanalytical Systems, Inc sells PAN membrane probes.

Semipermeable hollow fiber dialysis membranes can be obtained with a known molecular weight cut off (MWCO). The MWCO can be experimentally determined for a hollow fiber using several different experimental methods. The primary method used to determine MWCO for hollow fiber membranes is to continuously pass through the fiber lumen over a long period of time (24 h or greater) a solution containing known molecular weight markers. Known solutes that are rejected by the membrane are then used to calculate membrane MWCO. In practice, the MWCO is really not an absolute number, but rather the median of a range. This molecular weight rejection range is highly dependent on the semipermeable membrane materials pore distribution and can exhibit either a narrow or broad MWCO range (24).

Originally, the purpose of microdialysis sampling was to use the dialysis membrane as a means to provide a sample for chemical analysis that did not require further sample preparation steps such as protein removal. Intracranical dialysis applications typically target hydrophilic analytes with molecular weights < 500 Da. For these applications, dialysis membranes with low MWCO of $\sim$ 5000–6000 Da were commonly used to reject larger analytes and proteins so to allow liquid chromatographic analysis without further sample purification.

Recently, there has been a greater interest of applying microdialysis sampling to collect peptides and proteins. There have only been a few reports describing the use of different types of dialysis membranes towards the collection of large molecules, such as peptides and proteins (25). This is unfortunate as the types of commercially available membranes that are capable of providing the performance characteristics necessary for protein collection are relatively few. Kendrick extensively compared the recovery performance of different amino acids and peptides among different types of dialysis membranes (26). Torto *et al.* compared the dialysis collection efficiency for a series of saccharides [glucose (DP1), maltose (DP2), though maltoheptaose (D7)] among many different types of dialysis membranes (polyamide, polyethersulfone, and polysulfone) as well as different MWCO between 6 and 100 kDa (27). In some cases, membranes with similar MWCO and different chemistry exhibited similar recovery values. Whereas, some of the polysulfone membranes with 100 kDa MWCO exhibited quite low recovery for these low molecular weight analytes when compared to membranes with similar chemistry, but lower MWCO.

A set of model proteins including insulin (5.7 kDa), cytochrome *c* (12.4 kDa), ribonuclease A (13.7 kDa), lysozyme (14.4 kDa), and human serum albumin (67 kDa) were tested with different polymeric membranes and molecular weight cutoffs ranging between 20 and 150 kDa (28). All the membranes had similar external diameters (500 μm). Among the different membranes, only the polyethersulfone (100 kDa MWCO) commercially available from CMA Microdialysis, Inc and a Fresenius polysulfone membrane (150 kDa MWCO) exhibited similar recovery characteristics for the above-mentioned set of model proteins.

Membrane MWCO cannot be used as a means to specifically predict how well an analyte will be recovered during a microdialysis sampling procedures. New microdialysis sampling practitioners sometimes mistakenly believe that analytes near the membrane MWCO will be recovered. Although a membrane with 100 kDa MWCO allows some transport of molecules of this molecular weight, the recovery of an analyte of this size will be significantly $< 1\%$ (if at all) of the external sample concentration during microdialysis sampling. Dialysate analyte concentrations rarely reach equilibrium with the external sample concentrations except under unique conditions (very low flow rates and long membranes). For dialysate concentrations to reach those of the tissue medium surrounding the probe and thus approach equilibrium with the surrounding tissue concentrations, low perfusion fluid flow rates or long membranes are required in order to achieve residence times sufficient to obtain equilibration. During microdialysis sampling, the perfusion fluid only passes once through the inner membrane lumen with residence times on the order of seconds. Since this is in contrast to the methods used to obtain membrane MWCO, it is not surprising that sampled analyte molecular weight range is reduced due to the perfusion fluid making only one pass through the device. In general, the analyte molecular weight that easily passes through the membrane with 10% or greater recovery is roughly one tenth of the MWCO as shown in Fig. 2 (29). However, this is not an absolute value and different analytes have been reported to be difficult to dialyze across particular membranes. With rare exception (30), hydrophobic analytes typically are poorly dialyzed during microdialysis sampling (31,32). Furthermore, AN-69 membranes (polyacrylonitrile), which are sometimes used for kidney dialysis and have been used for in-house microdialysis probes, carry a negative charge that may cause rejection of certain negatively charged analytes (33,34).

Microdialysis sampling requires an inlet and outlet tube to be attached to the membrane. The length and inner diameter of the outlet tube attached to the membrane affects membrane backpressure. For some hollow fiber membranes, convective fluid loss (ultrafiltration) across

$$J_v = P(\Delta p - \Delta\pi)/l \tag{1}$$

these hollow fiber semipermeable membranes is possible and the extent of this ultrafiltration is related to the volumetric flux (J_v) shown in Eq. 1, where P is the permeability coefficient for the membrane, l is the length across the membrane (e.g., the membrane thickness), and Δp and $\Delta\pi$ the hydrostatic and osmotic pressure differences (35). Different membranes have difference permeability coefficients. The physical manifestation of this fluid loss is that the probe appears as if it is sweating during the dialysis procedure and is often observed with larger MWCO membranes.

Peptides and proteins diffuse very slowly across the small pores of membranes with MWCO between 5000 and 30,000 Da. To improve the relative recovery of these analytes, larger 100 kDa dialysis membranes have become commercially available. The disadvantage of these larger MWCO membranes is that they often exhibit ultrafiltration due to their larger pore sizes. Ultrafiltration fluid losses across 100 kDa or larger MWCO dialysis membranes should be determined prior to *in vivo* experiments. In particular, the ultrafiltration is exacerbated by the use of long outlet tubing with narrow diameter (a common need with awake and freely moving animal experiments). Osmotic balancing agents, such as dextrans or albumin, are commonly added to microdialysis perfusion fluids passed through 100 kDa or larger MWCO membranes to prevent excessive ultrafiltration as well as to prevent nonspecific adsorption on the device materials (36,37). While these agents are passed through the membrane, their potential loss to the surrounding tissue space has not been reported.

Recovery, Delivery, and Localized Infusion

Sensor devices are highly specific analytical detectors and can only be used to detect analytes that physically contact

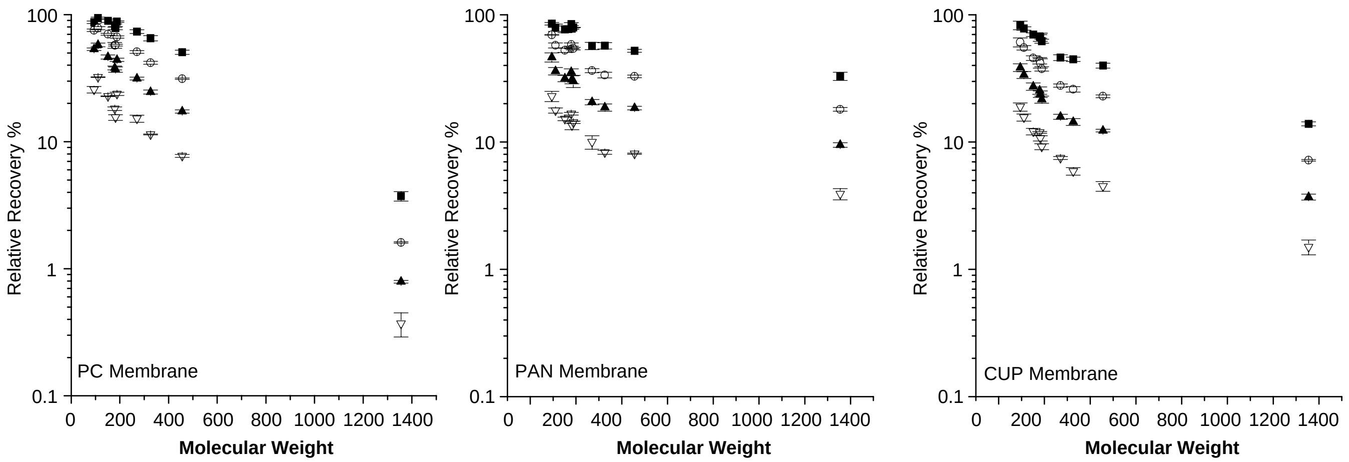


Figure 2. Semilog graph of relative recovery versus molecular weight for PC, PAN, and CUP membranes using different perfusion fluid flow rates. Flow rates are 0.5 μL·min^{-1} (■), 1.0 μL·min^{-1} (○), 2.0 μL·min^{-1} (▲) and 5.0 μL·min^{-1} (▲). Adapted from Ref. 29).

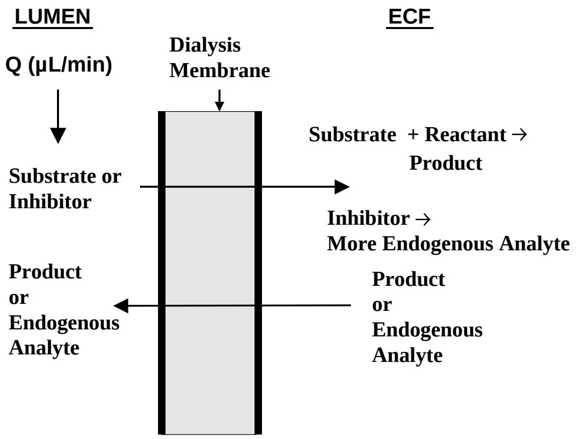

LUMEN
ECF
Q (μL/min)
Dialysis Membrane
Substrate + Reactant → Product
Substrate or Inhibitor
Inhibitor → More Endogenous Analyte
Product or Endogenous Analyte
Product or Endogenous Analyte

Figure 3. Localized infusion. A substrate or drug is locally infused through the microdialysis sampling probe. It diffuses into the ECF and then either reacts to form a product or causes a biochemical event to increase the concentrations of other molecules.

the sensor. Microdialysis sampling provides extensive flexibility with respect to ways in which it can be applied. Typical microdialysis sampling applications use what is termed the recovery mode where the device is placed into a sample matrix and analytes diffuse into the inner-fiber lumen of the probe (Fig. 1). Alternatively, the device can be used as a delivery device where a compound is infused through the probe causing either some alteration in a biochemical event (enzymatic reaction, enzymatic inhibition, or receptor binding) followed by collection of an endogenous analyte or enzymatic product. The probe can also be used to pass a substrate for a chemical or biochemical reaction and the products of that reaction can then be locally sampled as shown in Fig. 3. Unlike a specific sensor, microdialysis sampling devices allow for many physiological and biochemical processes to be studied within living tissue. Different examples of this approach for a variety of different tissues are shown in Table 3.

DEVICE CALIBRATION

Theoretical Foundations

Typical microdialysis sampling operating conditions (flow rates between 0.5 and 2.0 μL·min^{-1}) will yield a representative fraction of the analyte concentration in the surrounding ECF. Since most microdialysis conditions are such that equilibrium between the dialysate and the sample is not obtained, a calibration has to be used to relate dialysis concentrations to external sample concentrations. Extraction efficiency (E_d) is used to relate the dialysis concentration to the sample concentration. The steady-state E_d equation is

$$E_d = \frac{C_{\text{outlet}} - C_{\text{inlet}}}{C_{\text{tissue},\infty} - C_{\text{inlet}}} = 1 - \exp\frac{-1}{Q_{\text{d}}(R_{\text{d}} + R_{\text{m}} + R_{\text{e}} + R_{\text{t}})} \tag{2}$$

shown below in Eq. 2, where C_{outlet} is the analyte concentration exiting the microdialysis probe, C_{inlet} is the analyte concentration entering the microdialysis probe, $C_{\text{tissue},\infty}$ is the analyte tissue concentration far away from the probe, Q_{d} is the perfusion fluid flow rate and R_{d}, R_{m}, R_{e}, and R_{t} are a series of mass transport resistances for the dialysate, membrane, external sample, and a trauma layer that exists at the interface of the probe membrane and the tissue as defined in Scheme 1 (48,49).

The resistance terms are additive and understanding how these resistance terms affect E_{d} is vitally important with respect to experimental design. Throughout the microdialysis sampling process, collected analytes must diffuse through at least three regions (tissue, membrane and dialysate) in order to exit the microdialysis probe. Each mass transport resistance term defined in Scheme 1 has a diffusive component, which indicates that analytes with smaller diffusion coefficients will exhibit much lower E_{d}. The combined resistance contributions from R_{d} and R_{m} can be experimentally determined *in vitro* by collecting dialysates at different flow rates. A plot of the natural log of (1-E_{d}) versus 1/Q_{d} should yield a straight line, which can be regressed to determine the additive values for R_{d} and R_{m}. In addition to this information, an *in vitro* E_{d} experiment performed at 37 °C with stirring to cause the sample resistance, R_{e}, to approach a zero value, will yield the highest possible *in vivo* E_{d}.

The variables shown in Eq. 2 illustrate that a combination of perfusion fluid flow rate (Q_{d}), as well as mass transport resistances for the dialysate, membrane, and tissue medium external to the microdialysis probe affect E_{d}. Decreasing Q_{d} allows for a greater fluid residence time within the dialysis membrane thus allowing analyte concentration to increase along the membrane axis.

Table 3. Some Examples of Localized Infusion Using Microdialysis Sampling

Type (substrate or inhibitor)	Infused Compound	Measured Analyte or Application	Tissue	References
Inhibitor	Cocaine	Dopamine	Brain	38
Substrate	Substance P and other neuropeptides	Proteolytic Products	Brain	39,40
Substrate	Salicylic acid or 4-hydroxybenzoic acid	2,3-DHBA, 2,5-DHBA, 3,4-DHBA	*In Vitro*, Brain	41–43
Substrate	Phenol or acetaminophen	Metabolites	Liver	44,45
Substrate and inhibitor	Angiotensin, phosphoramidon, captopril	Metabolites and enzymatic inhibition	Renal Cortex	46
Substrate	Suc-(Ala)$_3$-pNA	Elastase (protease) activity	*In Vitro*	47

$$R_{\mathrm{d}} = \frac{13(r_{\mathrm{i}} - r_{\alpha})}{70\pi L r_{\mathrm{i}} D_{\mathrm{d}}} \; ; R_{\mathrm{m}} = \frac{\ln(r_{\mathrm{o}} / r_{\mathrm{i}})}{2\pi L D_{\mathrm{m}} \phi_{\mathrm{m}}} ; R_{\mathrm{e}} = \frac{\Gamma[K_{\mathrm{o}}(r_{\mathrm{o}} / \Gamma) / K_{1}(r_{\mathrm{o}} / \Gamma)]}{2\pi r_{\mathrm{o}} L D_{\mathrm{s}} \phi_{\mathrm{s}}}$$

$$\Gamma = \sqrt{\frac{D_{\mathrm{s}}}{(k_{\mathrm{ep}}(r) + k_{\mathrm{m}}(r) + k_{\mathrm{c}}(r))}}$$

Scheme 1. The multiple mass transport equations used to describe microdialysis sampling. D is the diffusion coefficient through the dialysate, D_{d}, membrane, D_{m}, and sample, D_{s}. The parameter L is the membrane length; Γ (cm) is a composite function; $k_{\mathrm{ep}}(r)$, $k_{\mathrm{m}}(r)$, and $k_{\mathrm{c}}(r)$ are kinetic rate constants as a function of radial position (r) from the microdialysis probe. Additional term definitions can be found in Ref. 48.

Fig. 4 shows a typical E_{d} curve simulated using the above equations (only R_{d} and R_{m} assuming a well-stirred system) for analytes with different aqueous diffusion coefficients. Fig. 4 clearly shows how microdialysis sampling membranes even for a hypothetical case perform in a manner that is consistent with diffusion being the major contributor affecting recovery. This scenario for the diffusivity is especially true for protein collection as many proteins of interest such as the cytokines have molecular weight values that can begin to approach the molecular weight cutoff limit for the dialysis membrane. In these cases, the protein diameter can begin to approach the values for the pore diameters resulting in restricted diffusion through the membrane, higher membrane mass transport resistances and thus reduced analyte recovery.

The parameter E_d is highly dependent on several physiochemical parameters (analyte diffusion coefficient, perfusion fluid flow rate, membrane pore size, and membrane surface area), kinetic uptake into cells (50) and the microvasculature (51), as well as the overall ECF volume fraction. The mass transport, resistances shown in Scheme 1 include analyte diffusivity terms for all three regions of mass transport, as well as kinetic terms for the tissue space, as shown in the above equations. Tissue diffusive and kinetic properties of the sample surrounding an implanted microdialysis probe will dictate the how reduced the *in vivo* E_{d} will be from the maximum possible *in vitro* E_{d} value at any particular flow rate. For hydrophilic analytes, it is generally assumed they diffuse only in the ECF that surrounds the tissue cellular components. This ECF space comprises approximately 20% of the overall tissue volume (52). Typically hydrophilic analytes have to diffuse around the cells en route to the microdialysis probe, the overall effective diffusive path length is increased due to the tortuous path traversed by the analyte. This tortuosity alters the tissue diffusion coefficient which can be approximated using $D_{\mathrm{ecf}} = D_{\mathrm{aq}}/\lambda^2$, where λ has a value of $\sim$ 1.5. In addition to the alteration in diffusive characteristics, tissues are vascularized and have active cellular components. Depending on the analyte, the active components will affect the overall microdialysis E_{d}.

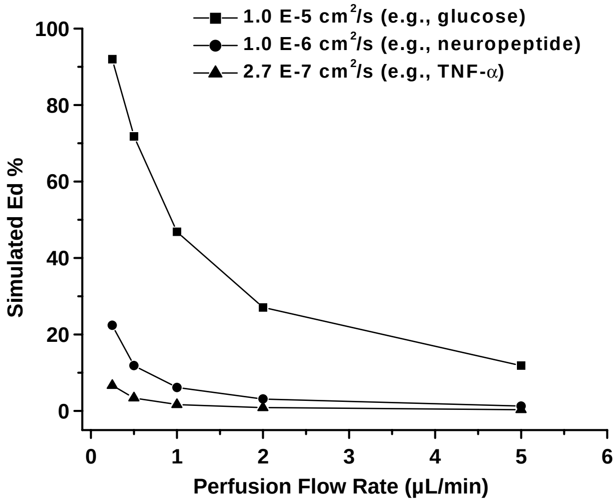


Figure 4. Simulated E_{d} using Bungay et al. (48) model with the following aqueous diffusion (D_{aq}) coefficients and with a membrane diffusion value of $0.2D_{\mathrm{aq}}$ with length of 10 mm and R_{i} (200 µm) and R_{o} (250 µm).

In addition to these parameters influencing microdialysis E_{d}, analyte properties affect the shape and the time to reach steady state for the concentration profile to the microdialysis probe. Analytes that diffuse rapidly and have a rapid supply to the tissue have narrow concentration profiles to the dialysis probe. Conversely, analytes that slowly diffuse and are not readily supplied to the tissue space will have concentration profiles to the dialysis probe that are not as steep. During microdialysis material is removed from the sampling site and the extent to which matter is removed is a function diffusive and kinetic parameters applied to that particular analyte, for example, how rapidly it the analyte replenished to the ECF from either capillaries (drugs) or cellular release processes. This has been a concern by others particularly as it relates to the understanding of dopamine transmission in the brain (53). However, dopamine is a special case and has rapid release and uptake kinetics in the ECF. For analytes that are poorly transported across the capillary space in the brain along with analytes that do not undergo significant uptake (e.g., drugs), their relative recoveries are generally lower than those with higher uptake/kinetic rates. It may appear to be counterintuitive to think that analytes with very rapid kinetic removal from the space surrounding the microdialysis probe have increased relative recovery. However, the higher removal rates cause the concentration profile to the dialysis probe to have a much greater gradient to the device as compared to a poorly removed analyte which would have a much shallower concentration profile to the device. For analytes with similar ability to diffuse through the membrane, that is, their membrane diffusion coefficients are nearly equal, the flux should be greater for the sharper concentration gradient thus causing greater relative recovery.

The theoretical foundations for microdialysis sampling during steady state operations derived by Bungay et al. have been widely used to corroborate many different *in vivo* experimental observations. In a series of papers focused on neurotransmitters, Justice's group has studied how uptake inhibition decreases microdialysis E_{d} 38,54,55. Stenken et al. 56 showed that since kinetic removal of targeted analyte may in some cases be additive, the inhibition of a particular cytochrome P450 isoform for phenacetin and antipyrine metabolism did not significantly alter E_d. This suggested that multiple kinetic components (capillary permeability plus metabolism) are important for analyte removal from liver tissue. Elmquist has shown that transporter inhibition in brain causes alterations in E_d for different drugs (57).

Calibration Methods

If the different variables shown in Eq. 2 and Scheme 1 are known prior to experimentation, then it is possible to predict the *in vivo* E_d. However, the difficulty with *in vivo* experiments is that obtaining values for variables is highly challenging. Furthermore, it is difficult to obtain an absolute value for C_{tissue} and an *in vivo* calibration requires that C_{tissue} (Eq. 2) be known. Even today, many experiments employing microdialysis sampling are semiquantitative and dialysate concentrations obtained are approximations of C_{tissue} at best. Since quite often ratios of dialysate analyte concentrations are obtained before and after some input (e.g., pharmacological or physical) with no attempt to measure the C_{tissue} during the experiment. To date, the *in vivo* calibration of implanted devices, including microdialysis probes, is an active area of research in the analytical chemistry and bioengineering communities and many authors have reviewed this subject (58,59). The principal difficulty with respect to obtaining a reliable device calibration is the inability to fully reproduce *in vitro* all the salient physiological features of tissue including permeation across capillaries and uptake processes (60,61).

Initial microdialysis sampling calibration focused on using an *in vitro* E_d calculation to estimate tissue analyte concentration. This approach gives a rough estimate of E_d and may provide an incorrect calibration factor. Furthermore, *in vitro* methods used for E_d measurement are affected by temperature (microdialysis sampling again is inherently a diffusion separation method), as well as sample stirring. A well-stirred buffer medium provides a mass transport external medium mass transport resistance (R_e) that approaches a value of zero. It is important to note that a quiescent medium does provide diffusional mass transport resistance and thus relative recoveries performed *in vitro* under stirred conditions will be different than those performed using quiescent conditions (48). How close a quiescently determined *in vitro* E_d is to the *in vivo* E_d would be wholly dependent upon the tissue kinetic properties for the targeted analyte. In other words, an analyte, such as dopamine, may exhibit higher *in vivo* E_d than *in vitro* quiescent E_d due to its extensive uptake kinetics causing a steeper concentration gradient to the dialysis probe as compared to the *in vitro* quiescent E_d measurement. Differences in the ability of the analyte to diffuse through the tissue space due to increased tortuosity and decreased volume fraction led to empirical methods that could be used to amend *in vitro* relative recovery calibration determinations (62). These methods focused on differences in tissue diffusion properties, but did not include the role of kinetic affects on microdialysis E_d causing significant errors for estimating *in vivo* values for C_{tissue}.

Jacobson et al. (63) were the first to try to create a more analyte-specific calibration procedure for microdialysis sampling. In their work, varying the perfusion fluid flow rates through the dialysis probe derived an analyte-specific membrane mass transport coefficient, K, shown below in Eq. 3, where A is the membrane surface area and Q_d is the dialysate volumetric flow rate. Eq. 3

$$\frac{C_{outlet}}{C_{tissue}} = 1 - \exp(-KA/Q_d) \quad (3)$$

is similar to Eq. 2, showing how the model of Bungay et al. incorporated previously known experimental results. In this case, the product (-*KA*) is related to the sum of the fraction of the mass transport resistance terms. Experimental results from this work immediately showed that understanding the underlying *in vivo* mechanisms affecting microdialysis E_d was more complicated than initially expected. These researchers found that different amino acids exhibited different *in vivo* mass transport coefficients. This was unexpected since the amino acids would be expected to have very similar diffusion coefficients due to their similar molecular weight. This data began to lead to the understanding that analyte properties (diffusion and kinetics) in the tissue play a major role with respect to microdialysis sampling calibration. An extension of calibration approach of Jacobson et al. is to pass the perfusion fluid through the dialysis probe so slowly that zero flow is approached and nearly 100% relative recovery as shown in Fig. 5. In this case, the goal is to calculate C_{sample} by attempting to reach an equilibrium state across the microdialysis membrane (64).

The most widely used calibration method for microdialysis sampling is based on knowing that diffusive flux should not occur across the dialysis membrane when the analyte concentration inside the perfusion fluid matches the concentration external to the microdialysis probe. This method was originally demonstrated by Lönnroth and has been called by a variety of names including Lönnroth plot,

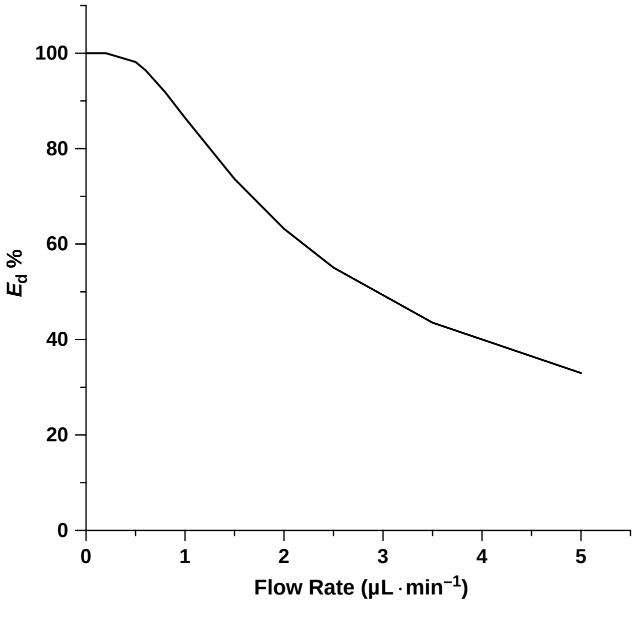


Figure 5. Mathematically modeled E_d for an approach to zero flow.

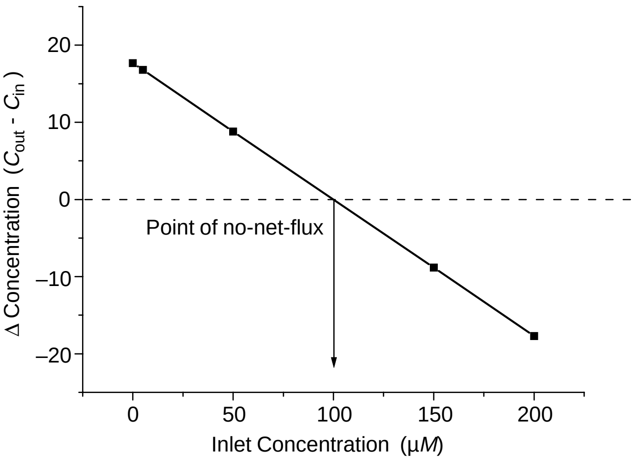


Figure 6. Lonnroth plot using hypothetical data.

method of no net flux (NNF) and method of zero net flux (ZNF). The ZNF method requires the tissue analyte concentration be at a steady state and has been used to determine basal concentrations for many analytes. The probe is perfused with different analyte concentrations that are either above or below the targeted analyte concentration. With these different perfusion studies, the loss or gain of analyte across the microdialysis probe can be determined and then plotted versus the inlet concentration as shown in Fig. 6. The analyte concentration is determined by the x-axis intercept and the relative recovery is the absolute value of the calibration line slope. A major drawback with this approach is the extensive amount of research time that has to be invested during the different perfusion studies. In particular, this is quite difficult to achieve with exogenous analytes (drugs) that would require a continuous infusion to achieve steady-state concentrations.

Quite often what is desired from a microdialysis sampling procedure is the analyte concentration during some sort of pharmacological challenge to an animal. This much needed analyte temporal information is not possible to collect with the requirement of steady state for the ZNF method. To overcome this problem and to gain information regarding the probe calibration from sample to sample, internal standards have been used during microdialysis sampling. Typically, internal standards have been chemicals with similar physicochemical properties as the targeted analyte. Some early work in internal standards proposed using one standard, such as antipyrine or ^{3}H–water to allow assessment of sample-to-sample differences should they arise throughout the duration of the microdialysis sampling events (65). Antipyrine would be a suitable reference for probe-to-probe variability because of its highly hydrophilic nature and ease of chemical detection. Urea has also been used as a microdialysis calibration reference for different metabolism studies (66–68). It is again important to note that the extraction efficiency of any particular analyte is a combined function of the different mass transport regions : dialysate, membrane, and most importantly tissue. Most likely during an *in vivo* microdialysis sampling experiment, the variability from sample to sample will occur due to alterations in tissue physiology, such as blood flow, metabolism, and uptake, which would serve to alter the tissue resistance and thus E_d. For this reason, it is generally preferred to have internal standards with similar tissue diffusion and kinetic properties as the analyte. Finding an appropriate internal standard is not a trivial task since analytes with similar structure and diffusive properties may also compete for enzymatic sites and may inhibit metabolic pathways that are important to removal and thus E_d values. However, this possibility must be considered in context of the tissue being sampled, as well as other additive kinetic properties (e.g., uptake into cells or capillaries) that may have a much greater impact on the E_d. For example, in the brain, the kinetic process that has been shown for several neurotransmitters to be most weighted towards affecting E_d are the neuronal uptake processes rather than metabolism processes. Additionally, in the liver, it appears that capillary blood flow and permeability are the primary contributors toward the E_d value obtained. Internal standards for peptides and proteins may be much harder to devise as receptor binding or for the cytokines, binding to the proteoglycan components of the ECF space may affect E_d and thus using molecular weight markers, such as inulin (69,70) or higher molecular weight fluorescein-labeled dextrans (e.g., FITC-Dextran 3000, FITC-Dextran 10,000) may only serve to report back diffusional mass transport differences during the duration of microdialysis sampling.

Effect of Probe Insertion Trauma

Insertion of microdialysis probes causes tissue damage (71,72). Although this has been known for quite some time, it has generally been overlooked by many microdialysis sampling users. The extent to which this insertion trauma affects the integrity of the microdialysis sampling concentrations and its true overall importance has been debated in the literature. The biomaterials literature is full of descriptions of the cellular events that occur after a foreign body implantation (73). It is known that edema occurs at the site of probe implantation (74) along with the recruitment of polymorphonuclear leukocytes (75,76) and matrix metalloproteinases (extracellular matrix remodeling enzymes) (77). Moderately reduced analyte flux to microdialysis probe chronically implanted has been reported for glucose (78).

The validity of the ZNF calibration methods for *in vivo* calibration, as well as determination of C_{tissue} for some analytes, has recently been a concern for neuroscientists interested in dopamine. Many careful studies performed by Michael's group illustrated that dopamine concentration measurements obtained with microelectrodes and microdialysis sampling devices were quite different (79,80). In particular, microdialysis sampling devices often exhibited much lower basal concentrations of dopamine than microelectrodes. Additional concerns have been raised for drug blood–brain barrier studies (81,82). Between these two examples, dopamine collection via microdialysis sampling appears to be the most severely affected because of its release and uptake sites being compromised due to the insertion trauma (49,83,84). In essence, the creation of a trauma layer creates four separate mass transport regions during microdialysis sampling : the dialysate, membrane, trauma layer, and normal tissue that need to be accounted for during data interpretation.

ANALYSIS OF MICRODIALYSIS SAMPLES

Microdialysis sampling is essentially married to appropriate detection methods for the collected dialysates. In addition to providing a means to sample from an *in vivo* site, microdialysis sampling also provides a relatively protein free or clean sample for chemical analysis. The only selectivity imparted into a microdialysis membrane is its molecular weight cutoff. For this reason, as long as a targeted analyte can diffuse through the membrane, the microdialysis sampling probe can be used as an *in vivo* chemical collection device. Thus, assuming the targeted analyte can pass through the dialysis membrane pores coupled with the appropriate analytical methods, a microdialysis sampling device could be considered to be essentially an all-purpose *in vivo* sensor (85). There is an extensive literature that has reviewed the associated analytical chemistry for making measurements in microdialysis samples (86). Additional reviews include: Adell et al. (87), Chaurasia (88), Church and Justice (89), Davies and Lunte (90), Horn (91), Kennedy (92), Kennedy et al. (93). Lunte et al. (94), Lunte and Lunte (95), Obrenovitch (96), Parkin et al. (97), and Parrot et al. (98).

Sample Volume Limitations

The major bottleneck 25 years ago for microdialysis sampling gaining more wide-spread and universal acceptance had to do with the analytical detection method sample volume limitations. During the early stages of microdialysis sampling, the primary analytical detection methods used for analyte quantification were liquid chromatography (LC) coupled with various types of detectors [ultraviolet–visible (UV–Vis), fluorescence, and electrochemical]. In addition to LC methods, radioimmunoassay (RIA) was occasionally used for peptides and proteins. Twenty-five years ago, it was not uncommon to require 25–50 μL of sample for LC analyses. Today, 50–100 μL of sample is still needed for standard immunoassays. The trade off that had to occur became one of either obtaining higher concentration recovery across the membrane by using low perfusion flow rates (1 $\mu L{\cdot}min^{-1}$ or less) or gaining sufficient temporal resolution by going to faster flow rates to achieve sufficient sample volumes for chemical analysis.

With the exception of glucose and lactate, many of the endogenous as well as xenobiotic analytes sampled using microdialysis had either micromolar (μM; 10^{-6} M) to nanomolar (nM 10^{-9} M) concentrations. These low concentrations often pushed the limitations of common analytical equipment since for most analytes an approximate detection limit with most UV-Vis detectors is roughly in the low μM range and for fluorescence and electrochemical detectors their detection limits are approximately in the nM range. The need to be able perform analytical measurements from such low volume dialysates drove analytical method development in multiple directions towards systems that could accommodate the low volumes without sacrificing method sensitivity, as well as development of high throughput methods that allowed for increased temporal resolution. Presently, there are many commercially available technologies that allow for samples that are <1 μL (e.g., capillary electrophoresis) or have duty cycles that are <1 min.

Separations-Based Methods for Microdialysis Sample Quantitation

Using separation methods, such as LC or capillary electrophoresis, for the quantitation of microdialysis samples is highly advantageous since these methods can be quickly adapted to many different analytes. Before the extensive use of microdialysis sampling for studies of neurochemical transmission, the use of *in vivo* voltammetry for analysis of neurotransmitters was just beginning to be described as a method for catecholamine (dopamine and norepinephrine) (99,100). The difficulty with using these methods was that electrode potentials needed to oxidize the catecholamines, as well as their metabolites (3, 4-dihydroxyphenylacetic acid, DOPAC) were similar. Furthermore, it was soon discovered that during vesicular release of dopamine, very high concentrations of ascorbic acid were released (101). For this reason, *in vivo* voltammetry of these important neurochemicals became more challenging since all of these chemicals can be oxidized at or below the same potential. The advantage of separations methods with appropriate detectors is that components including targeted analytes, as well as endogenous and exogenous interferences (see Fig. 7) can be appropriately separated and quantified. Thus, an additional advantage of using chromatographic methods is that chromatographic methods provide intrinsic multiplexing capabilities for the chemical analysis of microdialysis samples if different analytes are expected in the same samples.

Liquid Chromatography. Liquid chromatographic methods have been used for analyzing a broad class of analytes from microdialysis samples including catecholamines, amino acids, pharmaceuticals, and their metabolites. Several articles are available that describe the necessary requirements for microdialysis sample analysis using LC (102,103). Liquid chromatographic separations methods are well suited to microdialysis samples because of the high salt content contained in the perfusion fluids. Salts are generally not retained by the LC stationary phase and are therefore eluted in the chromatographic void volume. The resolving power of LC stationary phases allows for multiple analytes to be quantified during a single chromatographic run. Different detectors have been applied to LC separations for quantitation of dialysis samples.

Capillary Electrophoresis. Capillary electrophoresis (CE) is a separation method that involves passing an electric field across a micron-sized (∼25 to 75 μm internal diameter) capillary so as to allow separation of analytes based on their additive electrophoretic and electrosmotic mobilities. Neutral components in capillary electrophoresis

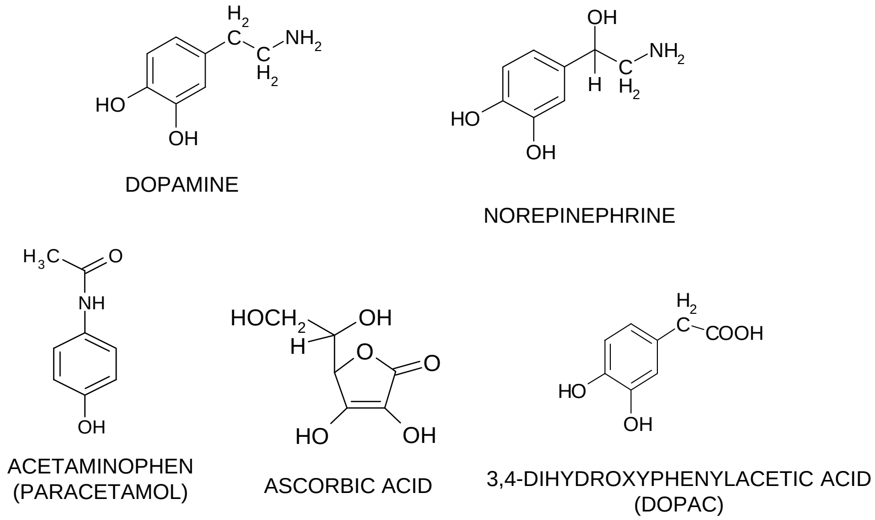

DOPAMINE
NOREPINEPHRINE
ACETAMINOPHEN (PARACETAMOL)
ASCORBIC ACID
3,4-DIHYDROXYPHENYLACETIC ACID (DOPAC)

Figure 7. Example of different analytes that can be oxidized at approximately the same potential using a carbon electrode under physiological conditions. The formal potential for dopamine, norepinephrine, and DOPAC are $\sim$ 0.7 V versus Ag/AgCl. For acetominophen, the formal potential is $\sim + 0.8$ V versus Ag/AgCl. For ascorbic acid, the oxidation formal potential is $\sim +$ 0.15 V versus Ag/AgCl.

will elute based on their electroosmotic mobility. Compared to LC analysis, CE provides greater separation efficiency and the possibility of faster separations. An additional advantage of CE is that an enormous research effort has been placed into microfabrication of CE devices onto microchips. This suggests the possibility for point-of-care technologies to be integrated with microdialysis sampling.

Like LC, CE has been used for a wide range of microdialysis sampling analyses including catecholamine neurotransmitters, amino acid neurotransmitters, and pharmaceutical compounds. Capillary electrophoresis has one main disadvantage and that has to do with the poor UV detection limits due to the path length being significantly reduced. However, sensitive detection can be achieved using electrochemical or laser-induced fluorescence detection approaches.

Although there are several disadvantages with CE detection, that is, no standard equipment, requirement of expertise, there is one advantage to this detection method for microdialysis samples. By using pH-stacking methods, significant on-column preconcentration (100 or greater) can be achieved (104). The mismatch of pH serves to greatly concentrate the sample zone on the head of the capillary column. This allows for online preconcentration to occur during the CE experiment. For collection and detection of peptides or proteins this preconcentration can serve to be highly useful. An additional advantage of CE is that extremely fast separations on the order of seconds can be achieved making this technology similar to that of a separations-based biosensor (105).

Fast Separations. Microdialysis sampling can be a slow temporal process due to the need to collect sufficient sample with sufficient relative recovery. To make microdialysis sampling more sensor like in terms of its response time, a number of groups have worked on achieving high speed separations, as well as direct coupling dialysate outflow the detection method. This is particularly important for studies of neurotransmitter dynamics. While electrochemical approaches for studies of catecholamine neurotransmission provide millisecond time resolution, microdialysis sampling is hindered by the turn-around time for the analytical method. High speed detection capability of cycle times of <1 min have been reported by many different groups using both liquid chromatographic, as well as capillary electrophoresis separations including 1 s time resolution with neurotransmitters (106) and <1 min resolution with pharmacokinetic analyses (107).

Examples of Different Types of Detection

Electrochemical Detection. Liquid chromatography coupled with electrochemical detection (LC–EC) is both a highly sensitive and selective method for analysis of compounds that can undergo an electrochemical reaction (oxidation or reduction). In this sense, LC–EC is highly suited to the chemical analysis of important biogenic amines (dopamine, norepinephrine, etc) obtained from microdialysis probes implanted into the brain (108). For these measurements, the electrochemical detector has a glassy carbon electrode that allows for oxidation of the amines at potentials of roughly 700 mV versus a Ag/AgCl reference electrode. The LC–EC excels in this analysis task because these analytes can be oxidized and furthermore their basal concentrations are in the low to mid-nanomolar range. The advantage of the separation is evident when considering that catecholamine metabolites (DOPAC, HVA) have basal concentrations that are $\sim$ 100–1000 times greater than the clinically relevant catecholamines (dopamine, norepinephrine, serotonin).

In addition to catecholamine detection of microdialysis samples, LC–EC analysis has been applied to low concentration analytes obtained under varying conditions of oxidative stress. In these cases, typically either salicylic acid or 4-hydroxybenzoic acid is directly infused through the microdialysis probe to locally deliver a trapping agent as shown in Fig. 3 (109). These benzoic acids react with hydroxyl radical to form catechols which can then be separated and detected by LC–EC (110). The LC–EC has also been used to quantify the DNA oxidative damage biomarker, 8-dOH–dGuanosine (111,112).

Electrochemical detection can be made to be more selective by altering the electrode surface. Gold electrodes coated with Hg to create an amalgam are highly selective toward thiols such as cysteine and glutathione with low potentials needed for oxidation ~ 150 mV versus Ag/AgCl (113). Lunte and O'Shea used this approach for glutathione detection using CE (114).

In addition to electrode modification, packed enzyme beds containing specific oxidase can be used prior to electrochemical detection. This is commonly applied to detection of the neurotransmitters choline and acetylcholine. Acetylcholine and choline can be separated chromatographically and then an enzymatic bed containing acetylcholine oxidase and choline oxidase is placed at the end of the column. These specific enzymatic reactions produce hydrogen peroxide which is then detected downstream at a platinum electrode (115). Note that more recent developments have attempted to immobilize the enzymes specifically to the electrode (116).

Examples of Fluorescence for Dialysates. Fluorescence detection is often employed when a known derivatization method can be applied to dialysate samples to improve method detection limits or to create a molecule that has a better handle for detection. For microdialysis samples, fluorescence derivatization is commonly applied to important amino acid neurotransmitters such as glutamate and GABA (117–119) and occasionally to biogenic amines (e.g., dopamine and norepinephrine) (120).

Examples of Mass Spectrometry for Dialysates. Mass spectrometry (MS) is a unique LC detector in that as long as the analyte has the ability to form an ion, MS can be used for analysis. However, mass spectrometric detection can be difficult with microdialysis samples because of the high salt content. This method has been particularly useful with neuropeptides because of their low concentrations. A problem with mass spectrometric detection is that salts from the dialysis perfusion fluid can cause analyte ionization suppression that leads to dramatically decreased detection capability for the method (121). Salts from dialysates are often removed via a column-switching technique that preconcentrates the analyte onto a C18 phase followed by desorption and detection (122). However, the use of nanoelectrospray devices can also reduce some of the problems associated with salt adducts (123–125). A particularly powerful method of LC-MS has been the ability to perform ionization in stages, which allows for structural elucidation of unknowns. The Kennedy group has been particularly successful with this approach for sequencing neuropeptides obtained from microdialysis samples (39,126).

Sensor Attachment to Microdialysis Probes

The microdialysis sampling process results in a relatively analytically clean (little to no protein) sample. Separations methods provide extensive analysis flexibility because they can be quickly optimized to the targeted analytes. However, there are *in vivo* monitoring situations where only one or a few analytes are targeted and highly specific analysis methods are available. A particular case in point is the continuous detection of glucose or lactate from diabetic humans (127). The primary advantage of coupling a sensing device to the end of the microdialysis sampling device is the sample matrix is simply saline passing across the analytical sensor. This prevents many of the difficulties associated with biofouling of implanted sensors (128). However, a critical problem for glucose sensing using this approach is that it can only provide information regarding the glucose concentration fluctuations throughout the day, but cannot really serve as an alarm system because of the lag times that are ~20–30 min as compared to normal glucose sensors of a few minutes (129). Despite this concern, there is great value in using specialized sensors to a microdialysis device because of the clean sample delivered to the sensing device.

The use of biosensors attached to the end of microdialysis probes has become highly useful for clinical neuroscience where measuring glucose and lactate and in some cases other neurotransmitters are needed to understand homeostatic mechanisms (130,131). Most biosensors attached to dialysis probes have been for glucose, lactate, or glutamate detection (132–134). Cook has published an interesting approach combining immunoassay with electrochemical detection for specific measurements of coritsol (135).

Immunoassay for Peptide and Protein Detection

Peptide and protein analysis of microdialysis samples is challenging since the concentrations of these targeted analytes are often in the ng·mL^{-1} to pg·mL^{-1} levels. This requires either highly sensitive fluorescence derivatization techniques for use with capillary electrophoresis (136) or sensitive immunoassays. Conventional immunoassays require 50–100 μL of sample. To obtain these volumes requires the use of high flow rates (2 μL·min^{-1} or greater) or very long collection times. In most cases, because highly sensitive radioimmunoassays (RIA) are used, higher flow rates are used to achieve moderate temporal resolution.

It is becoming increasingly evident that cellular communication in biological systems is highly complex and networked. Despite the tremendous growth in microdialysis sampling to monitor cellular biochemistry and an increased interest in peptide and protein detection in dialysates, there has been relatively little research towards new analytical methods that can detect peptides and proteins in low volume dialysate samples. A few approaches have been published that require 80–100 μL of sample for detection of several different proteins (137,128). Multiplexed assays (up to 25 analytes or more) that can be performed on a single sample have been recently created for immunology studies of the important inflammatory mediator class of cytokine proteins.

Highly sensitive multiplexed immunoassay platforms based on particle-based flow cytometry has become commercially available that allows cytokine measurements in 50-μL sample volumes (139,140). The limit of detection for these assays fall into the low pg/mL range comparable and have been compared and validated against standard ELISA methods (141). The use of these particle-based immunoassays is highly advantageous to the sample-limited microdialysis process and the sample volume needed has been decreased to < 25 μL by our group for cytokine detection. The advantage of these bead-based immunoassays for microdialysis samples is that several analytes can be analyzed in a single low volume sample. To illustrate the significant advantage that the bead-based immunoassay provides, if six separate cytokines were to be quantified in microdialysis samples using standard ELISA techniques more than 600 μL of sample would be needed. Using a flow rate of 1 $\mu\text{L}{\cdot}\text{min}^{-1}$, this would require 10 h of microdialysis sampling.

Mass versus Concentration Recovery

Microdialysis sampling E_d is a concentration recovery term and E_d increases as fluid flows decrease through the dialysis fiber creating longer residence times. Conversely, overall mass recovery typically increases as the flow rate increases as shown in Fig. 8. For some analytical applications, this increase in mass recovery may prove to be highly beneficial since it opens up possibilities for analytical preconcentration methods for the increased dialysate volumes.

MICRODIALYSIS SAMPLING APPLICATIONS

Microdialysis sampling applications have now been widely used in many different mammalian species including humans. The applications in humans have included studies in cancer, dermatology, immunology, pharmacokinetics and neuroscience. Many of these applications have been extensively reviewed by others and therefore will not be extensively discussed here. It is again important to note that microdialysis sampling has to date been principally applied to applications in neuroscience for the past three decades. However, over the past decade, more microdialysis sampling applications in other areas are now being described.

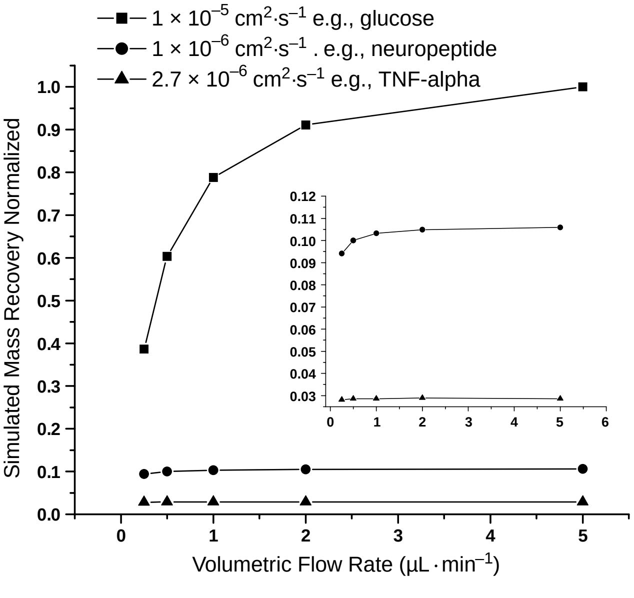


Figure 8. Modeled microdialysis sampling mass recovery for different analytes with different diffusion coefficients.

Neuroscience Applications

Microdialysis sampling has been in the neuroscientist toolbox for > 25 years. This device has been principally applied to neurotransmitter collection. Early on, the primary focus was in rat and more recently with probe redesigns and the biomedical value of knockouts, additional studies have been performed in mice (23,142). Current research interests focus on bridging the gap between animal models and human studies.

Reviewing all the microdialysis literature for neuroscience applications is a daunting task since the microdialysis sampling technique is now in wide use. Some of the applications have already been mentioned in the Analysis section of this article. However, several reviews and a book (see Bibliography section) are available as background. These reviews have covered general aspects of neurochemical collection with microdialysis sampling (143–146), microchemical analysis (147,148), and controversial aspects of neurotransmitter collection (84,149–151)

Neuropeptides. With successful sampling of hydrophilic neurotransmitters with microdialysis sampling, the next logical analyte class to target was neuropeptides. Like neurotransmitters, the quantitation of neuropeptides is challenging with microdialysis sampling due to their low concentrations. Furthermore, their lower diffusion coefficients cause their E_d values to be low. Temporal resolution can also be an issue since quite often immunoassays that require 50–100 μL are often used for detection. Despite these limitations, many neuropeptides have been sampled using microdialysis sampling and have been reviewed > 15 years ago (152,153). With the increased use of mass spectrometry for neuropeptide detection of dialysates (154,155) coupled with additional bead-based immunoassays, the application space for microdialysis sampling of neuropeptides should increase tremendously.

Pharmacokinetics

Microdialysis sampling has been applied for pharmacokinetic studies in animals and humans. The great advantage here is that microdialysis sampling tremendously decreases the overall number of animals used for a pharmacokinetics study. Typical pharmacokinetic studies in rodents require the animal to be sacrificed at each time point used for the analysis. Microdialysis sampling allows for collection throughout the time course of the experiment because it can be easily inserted into the jugular vein. Again, because of the highly flexible nature of liquid chromatographic analysis for drug studies, the microdialysis sampling technique can be rapidly applied to new drugs and their metabolites.

Microdialysis sampling applications in pharmacokinetics have been extensively reviewed. In addition to general reviews of the subject (156–158), there have been reviews focused on data analysis (159,160) and calibration (161,162). One of the more important points to consider when working with pharmacokinetic data obtained using microdialysis sampling is that a microdialysis sample represents a concentration average over the collection time period. This is in contrast to blood sampling that represents the analyte concentration directly at the sample time.

Clinical Applications

Clinical applications of microdialysis sampling have grown tremendously over the past decade and will continue to grow (163). At the present time, microdialysis sampling methods in human subjects have focused on studies in peripheral tissues for glucose collection (164–167) or drug distribution (168–170). Additional applications have been to determine gut barrier dysfunction (171). Microdialysis sampling applications in tumors has spanned both pharmacokinetic investigations (172–174), as well as collection of growth factors and cytokines (175). Microdialysis sampling has also been applied to human brain studies that have focused on understanding the underlying altered biochemistry that occurs when head trauma to drug distribution (176–180). Microdialysis sampling has been used as a means to monitor pharmacokinetics in the human dermis and has been compared to blister suction techniques (181). While both sampling methods produced similar data, it was found that microdialysis sampling was much easier to handle for both the patient and clinician.

Monitoring different growth factors and cytokines is becoming more important in clinical medicine since these proteins are known to affect cell-to-cell signaling and communication and are therefore becoming important biomarkers to measure. In particular, a group of proteins that are of great *in vivo* interest are the cytokines. Cytokines are potent, transient, and highly localized soluble messenger proteins (~6–80 kDa) produced by T-cells and macrophages that control nearly every aspect of the immune system (182). Cytokines exhibit complex interactions and therefore it is often more important to determine the concentration and cytokine profile after an immune challenge rather than the concentration of one single cytokine.

Microdialysis sampling is an ideal technique to achieve real time *in situ* monitoring of these important protein mediators and also has been recently described for proteomics applications (183). The application of microdialysis to this area is now emerging as potential approach for clinical *in vivo* studies in both healthy and diseased subjects to recover targeted cytokine molecules from the exact action sites and has recently been reviewed by Clough (25). Commercially available microdialysis probes with a 100 kDa MWCO membrane have been used for *in vivo* microdialysis of some cytokines (184,185). It is important to note that microdialysis sampling provides localized sampling and thus insight into localized concentrations of cytokines that cannot be achieved via sampling from blood plasma. This has been recently demonstrated with the cytokine IL-6 where its interstitial fluid concentration was 100-fold higher than that found in the plasma (186).

Cytokines have low E_d through 100 kDa membranes. To improve cytokine E_d larger MWCO membranes (3000 kDa) typically used for plasmaphoresis have been used (187). Others are beginning to use the 3000 kDa MWCO membrane for collection of IL-6 and TGF- β_1 (186,188–190). The range of *in vitro* recoveries for different cytokine proteins is shown in Table 4.

EVALUATION AND FUTURE USE

Microdialysis sampling has become a mature technology for neurotransmitter collection and pharmacokinetic determinations in animals. Clinical microdialysis sampling applications provide the greatest opportunity for growth. Despite the extensive biomedical use of conventional microdialysis sampling, there are still aspects of the device that could be tremendously improved.

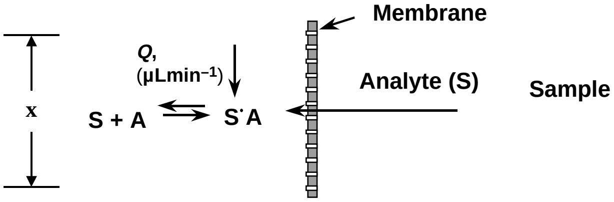
As currently practiced, microdialysis sampling in animals can be cumbersome due to the tubing lines required. Work in Lunte's group has focused on making micropumps using osmotic pumps as means to create line-free dialysis device (191,192). Reducing the microdialysis size by creating it on a microchip also has some advantages given that a decreased volume flow chamber may allow rapid equilibration across the device allowing E_d to approach nearly 100% (193–195).

A common problem with microdialysis sampling is the difficulty incurred when sampling hydrophobic analytes. This is an area with great promise with respect to either new device development or improvements to existing microdialysis sampling technology. Albumin is commonly included in the perfusion fluid to block

Table 4. Cytokine In Vitro Relative Recovery and Relevant Physicochemical Properties[a]

Cytokine	E_d% 0.5 μL·min^{-1}	E_d% 1.0 μL·min^{-1}	MW, kDa	Conformation	Active Protein, kDa
IL-2	4.5 ± 2.9 (12)	2.9 ± 1.1 (15)	17.2	Monomer	17.2
IL-4	12.0 ± 6.5 (12)	7.5 ± 2.6 (15)	13.6	Monomer	13.6
IL-5	1.3 ± 1.0 (12)	1.0 ± 0.5 (15)	13.1	Homodimer	26.2
IL-6	4.8 ± 1.4 (6)	2.7 ± 0.8 (6)	21.7	Monomer	21.7
IL-10	Not performed	1.1 ± 0.3 (3)	18.8	Homodimer	37.6
IL-12p70	N.D.	N.D.	35 & 40	Heterodimer	75
IFN-γ	2.0 ± 1.4 (18)	1.5 ± 0.8 (21)	15.9	Homodimer	31.8
MCP-1	24.5 ± 4.8 (6)	13.1 ± 3.9 (6)	13.1	Homodimer	26.2
TNF-α	8.0 ± 2.9 (15)	4.3 ± 1.1 (18)	17.3	Homotrimer	51.9

[a]Cytokine standards were either 1250 or 2500 pg·mL^{-1}. All solutions were quiescent at room temperature. A CMA/20 10-mm 100-kDa PES membrane was used for these studies performed in our laboratory. The numbers in parentheses after the RR values are the number of trials (n). All data are reported as mean ± SD. N.D. is not detected.



$$\left(\frac{dC_L}{dx}\right) = \frac{2pD_{\text{mem}}(C_o - C_L)}{Q\ln(R_o/R_i)} - \frac{pR_i^2kC_L}{Q}$$

$$\text{At } x = 0,\ C = C_o$$
$$\text{At } x = \text{L},\ C = C_\text{L}$$

Figure 9. Schematic of microdialysis enhanced transport. The analyte (S) or substrate for an affinity agent (A) binds with the affinity agent in the perfusion fluid. Affinity agents include antibodies, cyclodextrins or other supramolecular agents. The symbols in the equation are as follows: C_o = Concentration at $x = 0$, C_L = Concentration at $x = \text{L}$ (membrane length), D_mem = membrane diffusion coefficient, k = first-order rate constant for the reaction between the analyte, S, and the affinity agent, A, Q = perfusion fluid flow rate and R_i, R_o = inner and outer radii.

nonspecific adsorption sites within the microdialysis polymeric materials (196,197), and is still widely practiced (198). Unfortunately, this adds protein back into the dialysis perfusion that for small analytes causes difficulties with CE and LC analyses. A second approach involves using lipids, such as Intralipid, a material used to suspend hydrophobic drugs, to coat the nonspecific adsorption sites on the dialysis membrane or tubing (196,199). It is not known how compatible this approach is for CE or LC analysis.

Cyclodextrins have been used as perfusion-fluid additives for improving the microdialysis recovery for different analytes (200,201). A general scheme for the enhancement process is shown in Fig. 9. Cyclodextrins are well-known cyclic oligosacharides that have the capability to form inclusion complexes with various organic molecules by the capture of the guest molecule into a hydrophobic central cavity (202). Dialysate samples that contain cyclodextrin can be injected into an LC column without alteration to the chromatographic separation parameters, for example, plate number and peak width (203). Cyclodextrins are not selective, which for this enhancement approach is beneficial given that they can be applied to a wide variety of low molecular weight hydrophobic organic molecules (204,205). However, a difficulty with this approach is that cyclodextrins can diffuse out of the dialysis membrane that may complicate some applications. An additional difficulty is that cyclodextrins can interfere with affinity-based detection assays (immunoassays) due to competition between the analyte and cyclodextrin versus analyte and an antibody.

To overcome the free diffusion of cyclodextrin out of the dialysis tubing, different type of solid supports have been used for enhancement. Markides group has described the use of a solid-vehicle support for improving neuropeptide microdialysis relative recovery coupled with LC–MS detection (206,207). More specific enhancements can be achieved using antibody-immobilized beads for flow cytometry (208). With the antibody-enhancement approach, increases in microdialysis sampling relative recovery of 4–12-fold were achieved for different cytokines.

Microdialysis sampling is already beginning to be used in metabolomic studies using LC–MS detection methods (209). With the creation of microcoil nuclear magnetic resonance (NMR) that can measure nanoliter samples (210,211) the detection possibilities for dialysate samples are greatly increased. This approach has been recently applied to metabolomic studies with microdialysis sampling in brain (212).

The combination of these new discoveries for microdialysis sampling shows the enormous potential for solving many clinical biomedical problems with this device. The future for microdialysis sampling and device spin-offs is quite bright. New developments in instrumentation and collection procedures will provide much clinical benefit.

Microdialysis sampling has moved from a sampling device exclusively used to collect low molecular weight hydrophilic neurotransmitters to applications requiring the collection of larger proteins. Many cellular signaling processes occur via small molecule (nitric oxide, norepinephrine, acetylcholine, eicosanoids, etc.), peptide (angiotensin, etc.) as well as proteins (cytokines). Cytokine profiling for different disease states has become quite important. For the applications of both neuropeptide and larger protein collection such as the cytokines, the principal limitation to collecting these molecules is the diffusion properties of these analytes through the dialysis membrane. Proteins are difficult to collect through dialysis membranes since they can exhibit both nonspecific adsorption to the polymeric materials as well as diffusion restrictions through the membrane pores. Despite the variety of applications of microdialysis sampling to medical and scientific research, there have been few new developments with respect to improving the overall sampling efficiency for difficult to dialyze samples. As clinical proteomics and biomarker collection becomes better understood for making clinical predictions, it will be necessary to have sampling methods that can meet these clinical needs.

ACKNOWLEDGMENTS

Microdialysis sampling research in the Stenken laboratory is currently funded by NIH EB001441. The support from the John Wiley & Sons, Inc. editorial staff is greatly appreciated.

BIBLIOGRAPHY

1. Frangioni JV. *In vivo* near-infrared fluorescence imaging. Curr Opin Chem Biol 2003;7:626–634.
2. Ntziachristos V, Tung CH, Bremer C, Weissleder R. Fluorescence molecular tomography resolves protease activity *in vivo*. Nature Med 2002;8:757–761.
3. Sevick-Muraca EM, et al. Near-infrared imaging with fluorescent contrast agents. Handbook Biomed Fluoresc 2003; 445–527.
4. Myers RD, Adell A, Lankford MF. Simultaneous comparison of cerebral dialysis and push–pull perfusion in the brain of

rats: a critical review. Neurosci Biobehav Rev 1998;22:371–387.

5. Ellmerer M, et al. Measurement of interstitial albumin in human skeletal muscle and adipose tissue by open-flow microperfusion. Am J Physiol 2000;278:E352–E356.
6. Haaverstad R, Romslo I, Larsen S, Myhre HO. Protein concentration of subcutaneous interstitial fluid in the human leg. A comparison between the wick technique and the blister suction technique. Inter J Microcirculation, Clin Exper 1996;16:111–117.
7. Wiig H, Sibley L, DeCarlo M, Renkin EM. Sampling interstitial fluid from rat skeletal muscles by intermuscular wicks. Am J Physiol 1991;261:H155–H15.
8. Nedrebo T, et al. (2004). Differential cytokine response in interstitial fluid in skin and serum during experimental inflammation in rats. J Physiol 2004;556:193–202.
9. Bito LZ, et al. Concentration of free amino acids and other electrolytes in cerebrospinal fluid, in vivo dialyzate of brain, and blood plasma of the dog. J Neurochem 1966;13:1057–1067.
10. Delgado JM, et al. Dialytrode for long term intracerebral perfusion in awake monkeys. Arch Inter Pharmacodyn Ther 1972;198:9–21.
11. Ungerstedt U, Pycock C. Functional correlates of dopamine neurotransmission. Bull Schweizerischen Akad Medizinischen Wisse 1974;30:44–55.
12. Plock N, Kloft C. Microdialysis - theoretical background and recent implementation in applied life-sciences. Eur J Pharmaceut Sci 2005;25:1–24.
13. Davies MI, et al. Analytical considerations for microdialysis sampling. Adv Drug Deliv Rev 2000;45:169–188.
14. Benveniste H, Hüttemeier C. Microdialysis-theory and application. Prog Neurobiol 1990;35:195–215.
15. de Lange ECM, de Boer AG, Breimer DD. Methodological issues in microdialysis sampling for pharmacokinetic studies. Ad Drug Del Rev 2000;45:125–148.
16. Borg N, Stahle L. Recovery as a function of the osmolality of the perfusion medium in microdialysis experiments. Anal Chim Acta 1998;375:135–141.
17. Moghaddum B, Bunney BS. Ionic compositions of microdialysis perfusing solution alters the pharmacological responsiveness and basal outlflow of striatal dopamine. J Neurochem 1989;53:652–654.
18. Davies MI. A review of microdialysis sampling for pharmacokinetic applicaitons. Analyt Chim Acta 1999;379:227–249.
19. De Lange CMA, De Boer AG, Breimer DD. Intracerebral microdialysis. Intro Blood–Brain Barrier 1998; 94–112.
20. Bourne JA (2003). Intracerebral microdialysis: 30 years as a tool for the neuroscientist. Clin Exp Pharmacol Physiol 2003;30:16–24.
21. North American and International Commercial Sources of Microdialysis Probes. Available at CMA Microdialysis Inc. www.microdialysis.se; Bioanalytical Systems Inc., www. bioanalytical.com; SciPro www.scipro.com .
22. De la Peña A, Liu P, Derendorf H. Microdialysis in peripheral tissues. Adv Drug Deli Rev 2000;45:189–216.
23. Boschi G, Scherrmann JM. Microdialysis in mice for drug delivery research. Adv Drug Deliv Rev 2000;45:271–281.
24. Mulder M. Basic Principles of Membrane Technology. Dordrecht, The Netherlands: Kluwer Academic Publishers 1991.
25. Clough G. Microdialysis sampling of large molecules. The AAPS Journal E-Pub May 16, 2005.
26. Kendrick KM. Use of microdialysis in neuroendocrinology. Methods Enzymol 1989;168:182–205.
27. Torto N, et al. Optimal membrane choice for microdialysis sampling of oligosaccharides. J Chromatog A 1998;806:265–278.
28. Kjellstrom S, et al. Microdialysis—a membrane based sampling technique for quantitative determination of proteins. Chromatogr 1999;50:539–546.
29. Snyder KL, et al. Diffusion and calibration properties of microdialysis sampling membranes in biological media. Analyst 2001;126:1261–1268.
30. Schuck VJA, Rinas I, Derendorf H. *In vitro* microdialysis sampling of docetaxel. J Pharmaceutic Biomed Anal 2004;36: 807–813.
31. Groth L, Jírgensen A. *In vitro* microdialysis of hydrophilic and lipophilic compounds. Analyt Chim Acta 1997;355:75–83.
32. Mary S, et al. Assessment of the recovery of three lipophilic psoralens by microdialysis: An in vitro study. Inter J Pharmaceut 1998;161:7–13.
33. Patterson SL, Sluka KA, Arnold MA. A novel transverse push–pull microprobe: in vitro characterization and *in vivo* demonstration of the enzymatic production of adenosine in the spinal cord dorsal horn. J Neurochem 2001;76:234–246.
34. Patterson SL, Sluka KA, Arnold MA. A novel transverse push–pull microprobe: *In vitro* characterization and in vivo demonstration of the enzymatic production of adenosine in the spinal cord dorsal horn. [Erratum to document cited in A]. J Neurochem 2001;76:1955.
35. Ho WSW, Sikar KK, editors. Membrane Handbook New York; Van Nostrand Reinhold: 1992.
36. Rosdahl H, Hamrin K, Ungerstedt U, Henriksson J, A microdialysis method for the in situ investigation of the action of large peptide molecules in human skeletal muscle: Detection of local metabolic effects of insulin. Intern J Biol Macromole 2000;28:69–73.
37. Hamrin K, Rosdahl H, Ungerstedt U, Henriksson J. Microdialysis in human skeletal muscle: Effects of adding a colloid to the perfusate. J Appl Physiol 2002;92:385–393.
38. Smith AD, Justice JB. The effect of inhibition of synthesis, release, metabolism and uptake on the microdialysis extraction fraction of dopamine. J Neurosci Methods 1994;54:75–82.
39. Haskins WE, et al. Capillary LC-MS^2 at the attomole level for monitoring and discovering endogenous peptides in microdialysis samples collected *in vivo*. Analy Chem 2001;73:5005–5014.
40. Freed AL, Cooper JD, Davies MI, Lunte SM. Investigation of the metabolism of substance P in rat striatum by microdialysis sampling and capillary electrophoresis with laser-induced fluorescence detection. J Neurosc Methods 2001;109:23–29.
41. Ste-Mari L, Boismenu D, Vachon L, Montgomery J. Evaluation of sodium 4-hydroxybenzoate as an hydroxyl radical trap using gas chromatography-mass spectrometry and high-performance liquid chromatography with electrochemical detection. Ana Bioch 1996;241:67–74.
42. Chen R, Stenken JA. An in vitro hydroxyl radical generation assay for microdialysis sampling calibration. Anal Biochem 2002;306:40–49.
43. Marklund N, Clausen F, Lewander T, Hillered L. Monitoring of reactive oxygen species production after traumatic brain injury in rats with microdialysis and the 4-hydroxybenzoic acid trapping method. J Neurotrauma 2001;18: 1217–1227.
44. Scott DO, Lunte CE *In vivo* microdialysis sampling in the bile, blood, and liver of rats to study the disposition of phenol. Pharmaceut Res 1993;10:335–342.

45. Stenken JA, Ståthle L, Lunte CE, Southard MZ. Monitoring in situ liver metabolism in rats using microdialysis. Comparison of a microdialysis mass-transport model predictions to experimental metabolite generation data. J Pharmaceut Sci 1998;87:311–320.
46. Kajiro T, Nakajima Y, Fukushima T, Imai K. A method to evaluate the renin-angiotensin system in rat renal cortex using a microdialysis technique combined with HPLC-fluorescence detection. Anal Chem 2002;74:4519–4525.
47. Steuerwald AJ, Villeneuve JD, Sun L, Stenken JA. *In vitro* characterization of an in situ, microdialysis sampling assay for elastase activity detection. J Pharmaceut Biome Analy. In press.
48. Bungay PM, Morrison PF, Dedrick RL. Steady-state theory for quantitative microdialysis of solutes and water *in vivo* and in vitro. Life Sci 1990;46:105–119.
49. Bungay PM, et al. Microdialysis of dopamine interpreted with quantitative model incorporating probe implantation trauma. J Neurochem 2003;86:932–946.
50. Justice JB. Quantitative microdialysis of neurotransmitters. J Neurosc Methods 1993;48:263–276.
51. Clough GF, et al. Effects of blood flow on the *in vivo* recovery of a small diffusible molecule by microdialysis in human skin. J Pharmacol Exp Therapeut 2002;302:681–686.
52. Nicholson C. Diffusion and related transport mechanisms in brain tissue. Rep Prog Phys 2001;64:815–884.
53. Newton AP, Justice JB. Temporal response of microdialysis probes to local perfusion of dopamine and cocaine followed with one-minute sampling. Analy Chem 1994;66:1468–1472.
54. Cosford RJO, Vinson AP, Kukoyi S, Justice JBJ. Quantitative microdialysis of serotonin and norepinephrine: Pharmacological influences on *in vivo* extraction fraction. J Neurosci Methods 1996;68:39–47.
55. Vinson PN, Justice JB. Effect of neostigmine on concentration and extraction fraction of acetylcholine using quantitative microdialysis. J Neurosci Methods 1997;73:61–67.
56. Stenken JA, Lunte CE, Southard MZ, Ståhle L. Factors that influence microdialysis recovery. Comparison of experimental and theoretical microdialysis recoveries in rat liver. J Pharmaceut Sci 1997;86:958–966.
57. Dai H, Elmquist WF. Drug transport studies using quantitative microdialysis. Methods Mol Med 2003;89:249–264.
58. Stenken JA. Methods and issues in microdialysis calibration. Anal Chim Acta 1999;379:337–357.
59. Chen KC, et al. Theory relating in vitro and in vivo microdialysis with one or two probes. J Neurochem 2002;81:108–121.
60. Rice ME, Nicholson C. Diffusion and ion shifts in the brain extracellular microenvironment and their relevance for voltammetric measurements. The brain is not a beaker: *In vivo* vs. *In vitro* voltammetry. In: Boulton A, Baker G, Adams RN, editors. Neuromethods, Vol. 27, Voltammetric Methods in Brain Systems, ed. New York: Humana Press Inc.; 1995. pp 27–79.
61. Reach G, Wilson GS. Can continuous glucose monitoring be used for the treatment of diabetes? Analy Chem 1992;64: 381A–386A.
62. Benveniste H, Hüttemeir C. Microdialysis-theory and application. Progr Neurobiol 1990;35:195–215.
63. Jacobson I, Sandberg M, Hamberger A. Mass transfer in brain dialysis devices—a new method for the estimation of extracellular amino acids concentration. J Neurosci Methods 1985;15: 263–268.
64. Menacherry S, Hubert W, Justice JB. *In vivo* calibration of microdialysis probes for exogenous compounds. Analyt Chem 1992;64:577–583.
65. Yokel RA, Allen DD, Burgio DE. McNamara JP. Antipyrine as a dialyzable reference to correct differences in efficiency among and within sampling devices during in vivo microdialysis. J Pharmacol Toxicol Methods 1992;27:135–142.
66. Strindberg L, Lonnroth P. Validation of an endogenous reference technique for the calibration of microdialysis catheters. Scand J Clin Lab Investigation 2000;60:205–212.
67. Ronne-Engstrom E, et al. Intracerebral microdialysis in neurointensive care: The use of urea as an endogenous reference compound. J Neurosur 2001;94:397–402.
68. Ettinger SN, et al. Urea as a recovery marker for quantitative assessment of tumor interstitial solutes with microdialysis. Cancer Res 2001;61:7964–7970.
69. Leypoldt JK, Burkart JM. Small-solute and middle-molecule clearances during continuous flow peritoneal dialysis. Adv Peritoneal Dialysis 2002;18:26–31.
70. Krejcie TC, et al. Modifications of blood volume alter the disposition of markers of blood volume, extracellular fluid, and total body water. J Pharmacol Exp Therap 1999;291: 1308–1316.
71. Benveniste H, Diemer NH. Cellular reactions to implantation of a microdialysis tube in the rat hippocampus. Acta Neuropathol 1987;74:234–238.
72. Grabb MC, et al. Neurochemical and morphological responses to acutely and chronically implanted brain microdialysis probes. J Neurosci Methods 1998;82:25–34.
73. Anderson JM. Biological responses to materials. Ann Rev Mater Res 2001;31:81–110.
74. Dykstra KH, et al. Quantitative examination of tissue concentration profiles associated with microdialysis. J Neurochem 1992;58:931–940.
75. Davies MI, Lunte CE. Microdialysis sampling for hepatic metabolism studies: impact of microdialysis probe design and implantation technique on liver tissue. Drug Metabolism Disposition 1995;23:1072–1079.
76. Clapp-Lilly KL, et al. An ultrastructural analysis of tissue surrounding a microdialysis probe. J Neurosci Methods, 1999;90:129–142.
77. Planas AM, et al. Certain forms of matrix metalloproteinase-9 accumulate in the extracellular space after microdialysis probe implantation and middle cerebral artery occlusion/reperfusion. J Cerebral Blood Flow Metabolism 2002;22: 918–925.
78. Wisniewski N, et al. Analyte flux through chronically implanted subcutaneous polyamide membranes differs in humans and rats. Am J Physiol, Endocrinol Metabolism 2002;282:E1316–E1323.
79. Lu Y, Peters JL, Michael AC. Direct comparison of the response of voltammetry and microdialysis to electrically evoked release of striatal dopamine. J Neurochem 1998;70:584–593.
80. Borland LM, Shi G, Yang H, Michael AC. Voltammetric study of extracellular dopamine near microdialysis probes acutely implanted in the striatum of the anesthetized rat. J Neurosci Methods 2005;146:149–158.
81. Morgan ME, Singhal D, Anderson BD. Quantitative assessment of blood-brain barrier damage during microdialysis. J Pharmacol Exp Therap 1996;277:1167–1176.
82. Groothuis DR, et al. Changes in blood-brain barrier permeability associated with insertion of brain cannulas and microdialysis probes. Brain Res 1998;803:218–230.
83. Chen KC. Preferentially impaired neurotransmitter release sites not their discreteness compromise the validity of microdialysis zero-net-flux method. J Neurochem 2005;92: 29–45.

84. Khan SA, Michael AC. Invasive consequences of using microelectrodes and microdialysis probes in the brain. TrAC, Trends Anal Chem 2003;22:503–508.
85. Ballerstadt R, Schultz JS. Sensor methods for use with microdialysis and ultrafiltration. Adv Drug Del Rev 1996; 21:225–238.
86. Davies MI, et al. Analytical considerations for microdialysis sampling. Adv Drug Del Rev 2000;45:169–188.
87. Adell A, Artigas F. *In vivo* brain microdialysis: Principles and applications. Neuromethods 1998;32:1–33.
88. Chaurasia CS. *In vivo* microdialysis sampling: Theory and applications. Biomed Chromatog 1999;13:317–332.
89. Church WH, Justice Jr JB. On-line small-bore chromatography for neurochemical analysis in the brain. Adv Chromatog 1989;28:165–194.
90. Davies IM, Lunte CE. Microdialysis sampling coupled on-line to microseparation techniques. Chem Soc Rev 1997;26:215–222.
91. Horn TFW, Engelmann M. *In vivo* microdialysis for nonapeptides in rat brain—a practical guide. Methods 2001;23:41–53.
92. Kennedy RT. Bioanalytical applications of fast capillary electrophoresis. Anal Chim Acta 1999;400:163–180.
93. Kennedy RT, et al. *In vivo* neurochemical monitoring by microdialysis and capillary separations. Curr Opin Chem Biol 2002;6:659–665.
94. Lunte CE, Scott DO, Kissinger PT. Sampling living systems using microdialysis probes. Anal Chem, 1991;63:773A–780A.
95. Lunte SM, Lunte CE. Microdialysis sampling for pharmacological studies: HPLC and CE analysis. In: Brown PR, Grushka E, editors. Advances in Chromatography. New York: Marcel Dekker; 1996. pp 383–432.
96. Obrenovitch TP, Zilkha E. Microdialysis coupled to online enzymatic assays. Methods 2001;23:63–71.
97. Parkin MC, Hopwood SE, Boutelle MG, Strong AJ. Resolving dynamic changes in brain metabolism using biosensors and on-line microdialysis. TrAC, Trends Analy Chem 2003;22:487–497.
98. Parrot S, et al. Microdialysis monitoring of catecholamines and excitatory amino acids in the rat and mouse brain: Recent developments based on capillary electrophoresis with laser-induced fluorescence detection-A mini-review. Cellular Mol Neurobiol 2003;23:793–804.
99. Wightman RM, Strope E, Plotsky PM, Adams RN. Monitoring of transmitter metabolites by voltammetry in cerebrospinal fluid following neural path stimulation. Nature London 1976;262:145–146.
100. Adams RN. Probing brain chemistry with electroanalytical techniques. Anal Chem 1976;48:1126A–1138A.
101. Nagy G, Rice ME, Adams RN. A new type of enzyme electrode: The ascorbic acid eliminator electrode. Life Sci 1982; 31:2611–2616.
102. Wages SA, Church WH, Justice Jr JB. Sampling considerations for on-line microbore liquid chromatography of brain dialysate. Analy Chem 1986;58:1649–1656.
103. Kennedy RT, German I, Thompson JE, Witowski SR. Fast analytical-scale separations by capillary electrophoresis and liquid chromatography. Chem Rev 1999;99:3081–3131.
104. Arnett SD, Lunte CE. Investigation of the mechanism of pH-mediated stacking of anions for the analysis of physiological samples by capillary electrophoresis. Electrophoresis 2003;24:1745–1752.
105. Davies M, et al. Studies on animal to instrument hyphenation: Development of separation-based sensors for near real-time monitoring of drugs and neurotransmitters. Sample Preparation for Hyphenated Analyt Tech 2004; 191–220.
106. Rossell S, Gonzalez LE, Hernandez L. One-second time resolution brain microdialysis in fully awake rats. Protocol for the collection, separation and sorting of nanoliter dialyzate volumes. J Chromatogr B: Analy Technol Biomed Life Sci 2003;784:385–393.
107. Chen A, Lunte CE. Microdialysis sampling coupled on-line to fast microbore chromatography. J Chromatog Analy Appl 1995;691:29–35.
108. Kissinger PT, Refshauge C, Dreiling R, Adams RN. Electrochemical detector for liquid chromatography with picogram sensitivity. Analyt Lett 1973;6:465–477.
109. Acworth IN, Bogdanov MB, McCabe DR, Beal MF. Estimation of hydroxyl free radical levels *in vivo* based on liquid chromatography with electrochemical detection. Methods Enzymol 1999;300:297–313.
110. Acworth IN, Bailey BA, Maher TJ. The use of HPLC with electrochemical detection to monitor reactive oxygen and nitrogen species, markers of oxidative damage and antioxidants: application to the neurosciences. Prog HPLC-HPCE 1998;7:3–56.
111. Bogdanov MB, et al. A carbon column-based liquid chromatography electrochemical approach to routine 8-hydroxy-2’-deoxyguanosine measurements in urine and other biologic matrices: A one-year evaluation of methods. Free Rad Biol Med 1999;27:647–666.
112. Bogdanov MB, et al. Increased oxidative damage to DNA in a transgenic mouse model of Huntington’s disease. J Neurochem 2001;79:1246–1249.
113. Stenken JA, Puckett DL, Lunte SM, Lunte CE. Detection of *N*-acetylcysteine, cysteine and their disulfides in urine by liquid chromatography with a dual-electrode amperometric detector. J Pharmaceut Biomed Analysis 1990;8:85–89.
114. Lunte SM, O’Shea TJ. Pharmaceutical and biomedical applications of capillary electrophoresis/electrochemistry. Electrophoresis 1994;15:79–86.
115. Zackheim JA, Abercrombie ED. HPLC/EC detection and quantification of acetylcholine in dialysates. Methods Mol Med 2003;79:433–441.
116. Kehr J, Dechent P, Kato T, Ogren SO. Simultaneous determination of acetylcholine, choline and physostigmine in microdialysis samples from rat hippocampus by microbore liquid chromatography/electrochemistry on peroxidase redox polymer coated electrodes. J Neurosci Methods 1998;83:143–150.
117. Kehr J. Determination of glutamate and aspartate in microdialysis samples by reversed-phase column liquid chromatography with fluorescence and electrochemical detection. J Chromatog B: Biomed Sci App 1998;708:27–38.
118. Kehr J. Determination of g-aminobutyric acid in microdialysis samples by microbore column liquid chromatography and fluorescence detection. J Chromatog B: Biomed Sci App 1998;708:49–54.
119. Tao L, Thompson JE, Kennedy RT. Optically gated capillary electrophoresis of o-phthalaldehyde/b-mercaptoethanol derivatives of amino acids for chemical monitoring. Analy Chem 1998;70:4015–4022.
120. Yamaguchi M, et al. Determination of norepinephrine in microdialysis samples by microbore column liquid chromatography with fluorescence detection following derivatization with benzylamine. Analy Biochem 1999;270: 296–302.
121. Mann M, Fenn JB. Electrospray mass spectrometry: principles and methods. Mass Spectrom 1992;1:1–35.
122. Emmett MR, Andren PE, Caprioli RM. Specific molecular mass detection of endogenously released neuropeptides using *in vivo* microdialysis/mass spectrometry. J Neurosci Methods 1995;62:141–147.

123. Wilm M, Mann M. Analytical properties of the nanoelectrospray ion source. Analy Chem 1996;68:1–8.
124. Juraschek R, Dulcks T, Karas M. Nanoelectrospray-more than just a minimized-flow electrospray ionization source. J Am Soc Mass Spectrom 1999;10:300–308.
125. Gangl ET, Annan M, Spooner N, Vouros P. Reduction of signal suppression effects in ESI-MS using a nanosplitting device. Analy Chem 2001;73:5635–5644.
126. Haskins WE, et al. Discovery and neurochemical screening of peptides in brain extracellular fluid by chemical analysis of *in vivo* microdialysis samples. Analy Chem 2004;76:5523–5533.
127. Heinemann L. Continuous glucose monitoring by means of the microdialysis technique: Underlying fundamental aspects. Diabetes Technol Therap 2003;5:545–561.
128. Wisniewski N, Moussy F, Reichert WM. Characterization of implantable biosensor membrane fouling. Fresenius J Analy Chem 2000;366:611–621.
129. Ward WK. Subcutaneous glucose monitoring: Microdialysis vs. intracorporeal. Diabetes Care 2002;25:410–411.
130. O'Neill RD, Lowry JP, Mas M. Monitoring brain chemistry *in vivo*: Voltammetric techniques, sensors, and behavioral applications. Crit Rev Neurobiol 1998;12:69–127.
131. Jones DA, et al. On-line monitoring in neurointensive care Enzyme-based electrochemical assay for simultaneous, continuous monitoring of glucose and lactate from critical care patients. J Electroanalyt Chem, 2002; 538–539, 243–252.
132. Boutelle MG, Fellows LK, Cook C. Enzyme packed bed system for the on-line measurement of glucose, glutamate, and lactate in brain microdialysis. Analy Chem 1992;64: 1790–1794.
133. Volpe G, Moscone D, Compagnone D, Palleschi G. *In vivo* continuous monitoring of ***L-lactate coupling subcutaneous microdialysis and an electrochemical biocell. Sensors Actuators B 1995; 24–25 138–141.
134. Ryan MR, Lowry JP, O'Neill RD. Biosensor for neurotransmitter L-glutamic acid designed for efficient use of L-glutamate oxidase and effective rejection of interference. Analyst 1997;122:1419–1424.
135. Cook CJ. Real-time measurement of corticosteroids in conscious animals using an antibody-based electrode. Nature Biotechnol 1997;15:467–471.
136. Sandberg M, Weber SG. Techniques for neuropeptide determination. TrAC, Trends Analy Chem 2003;22:522–527.
137. O'Connor KA, et al. A method for measuring multiple cytokines from small samples. Brain, Behavior, Immunity 2004; 18:274–280.
138. Li Y, Schutte RJ, Abu-Shakra A, Reichert WM. Protein array method for assessing *in vitro* biomaterial-induced cytokine expression. Biomaterials 2005;26:1081–1085.
139. Vignali DAA. Multiplexed particle-based flow cytometric assays. J Immunol Methods 2000;243:243–255.
140. Kellar KL, et al. Multiplexed fluorescent bead-based immunoassays for quantitation of human cytokines in serum and culture supernatants. Cytometry 2001;45:27–36.
141. Kellar KL, Iannone MA. Multiplexed microsphere-based flow cytometric assays. Exp Hemato 2002;30:1227–1237.
142. Xie R, Hammarlund-Udenaes M, De Boer AG, De Lange ECM. The role of P-glycoprotein in blood-brain barrier transport of morphine: Transcortical microdialysis studies in mdr1a (−/−) and mdr1a (+/+) mice. Br J Pharmacol 1999; 128:563–568.
143. Ungerstedt U. Introduction to intracerebral microdialysis. Tech Behav Neural Sci 1991;7:3–22.
144. Sharp T, Zetterstrom T. *In vivo* measurement of monoamine neurotransmitter release using brain microdialysis. Monit. Neuronal Act 1992; 147–179.
145. Westerink BHC, Timmerman W. Do neurotransmitters sampled by brain microdialysis reflect functional release? Anal Chim Acta 1999;379:263–274.
146. Salamone JD. The behavioral neurochem of motivation: methodological and conceptual issues in studies of the dynamic activity of nucleus accumbens dopamine. J Neurosc Methods 1996;64:137–149.
147. Justice Jr JB. Microchemical analysis in the brain. Microchem J 1986;34:11–14.
148. Kennedy RT, et al. *In vivo* neurochemical monitoring by microdialysis and capillary separations. Curr Opin Chem Biol 2002;6:659–665.
149. Fuxe K, Ferre S, Zoli M, Agnati LF. Integrated events in central dopamine transmission as analyzed at multiple levels. Evidence for intramembrane adenosine A2A/dopamine D2 and adenosine A1/dopamine D1 receptor interactions in the basal ganglia. Brain Res Rev 1998;26:258–273.
150. Del Arco A, Segovia G, Fuxe K, Mora F. Changes in dialysate concentrations of glutamate and GABA in the brain: An index of volume transmission mediated actions? J Neurochem 2003;85:23–33.
151. Di Chiara G, Tanda G, Carboni E. Estimation of in-vivo neurotransmitter release by brain microdialysis: The issue of validity. Behav Pharmacol 1996;7:640–657.
152. Kendrick KM. Use of microdialysis in neuroendocrinology. Methods Enzymol 1989;168:182–205.
153. Kendrick KM. Microdialysis measurement of *in vivo* neuropeptide release. J Neurosc Methods 1990;34:35–46.
154. Andren PE, Lin-S. N, Caprioli RM. Microdialysis/mass spectrometry. Mass Spectrom 1994;2:237–254.
155. Andren PE, Farmer TB, Klintenberg R. Endogenous release and metabolism of neuropeptides utilizing *in vivo* microdialysis microelectrospray mass spectrometry. Mass Spectrom Hyphenated Tech Neuropeptide Res 2002; 193–213.
156. Ståhle L, Microdialysis in pharmacokinetics. Eur J Drug Metabol Pharmacokine 1993;18:89–96.
157. de Lange ECM, Danhof M, de Boer AG. Breimer DD. Methodological considerations of intracerebral microdialysis in pharmacokinetic studies on drug transport across the blood-brain barrier. Brain Res Rev 1997;25:27–49.
158. Hansen DK, et al. Pharmacokinetic and metabolism studies using microdialysis sampling. J Pharmaceut Sci 1999;88: 14–27.
159. Ståhle L. Pharmacokinetic estimations from microdialysis data. Eur J Clin Pharmacol 1992;43:289–294.
160. Ståhle L. Zero and first moment area estimation from microdialysis data. Eur J Clin Pharmacol 1993;45:477–481.
161. Hammarlund-Udenaes M, Paalzow LK. de Lange ECM. Drug equilibration across the blood-brain barrier—pharmacokinetic considerations based on the microdialysis method. Pharmaceut Res 1997;14:128–134.
162. Bungay PM, Dedrick RL, Fox E, Balis FM. Probe calibration in transient microdialysis *in vivo*. Pharmaceut Res 2001;18: 361–366.
163. Muller M. Science, medicine, and the future: Microdialysis. Br Med J 2000;324:588–591.
164. De Boer J, Korf J, Plijter-Groendijk H. *In vivo* monitoring of lactate and glucose with microdialysis and enzyme reactors in intensive care medicine. Inter J Artif Organs 1994;17:163–170.
165. Weintjes KJ, et al. Microdialysis of glucose in subcutaneous adipose tissue up to 3 weeks in healthy volunteers. Diabetes Care 1998;21:1481–1488.
166. Gudbjoernsdottir S, et al. Direct measurements of the permeability surface area for insulin and glucose in human

skeletal muscle. J Clin Endocrino Metab 2003;88:4559–4564.

167. Lonnroth P. Microdialysis in adipose tissue and skeletal muscle. Hormone Metabol Res 1997;29:344–346.

168. Blochl-Daum B, et al. Measurement of extracellular fluid carboplatin kinetics in melanoma metastases with microdialysis. Br J Cancer 1996;73:920–924.

169. Muller M, et al. *In vivo* drug-response measurements in target tissues by microdialysis. Clin Pharmacol Therapeut 1997;62:165–170.

170. Lindberger M, Tomson T. Ståhle L. Validation of microdialysis sampling for subcutaneous extracellular valproic acid in humans. Therapeut Drug Monitoring 1998;20:358–362.

171. Solligård E, et al. Gut barrier dysfunction as detected by intestinal luminal microdialysis. Int Care Med 2004;30: 1188–1194.

172. Chu J, Gallo JM. Application of microdialysis to characterize drug disposition in tumors. Adv Drug Del Rev 2000;15:243–253.

173. Mader RM, et al. Penetration of capecitabine and its metabolites into malignant and healthy tissues of patients with advanced breast cancer. Br J Cancer 2003;88:782–787.

174. Johansen MJ, Thapar N, Newman RA, Madden T. Use of microdialysis to study platinum anticancer agent pharmacokinetics in preclinical models. J Expe Therap Oncol 2002; 2:163–173.

175. Dabrosin C. Microdialysis — an *in vivo* technique for studies of growth factors in breast cancer. Frontiers Biosci 2005;10: 1329–1335.

176. Hamani C, Luer MS, Dujovny M. Microdialysis in the human brain: review of its applications. Neurolog Res 1997;19:281–288.

177. Vespa P, et al. Increase in extracellular glutamate caused by reduced cerebral perfusion pressure and seizures after human traumatic brain injury: A microdialysis study. J Neurosur 1998;89:971–982.

178. Sherwin AL. Neuroactive amino acids in focally epileptic human brain: A review. Neurochem Res 1999;24:1385–1395.

179. Bradberry CW. Applications of microdialysis methodology in nonhuman primates: Practice and rationale. Crit Rev Neurobiol 2000;14:143–163.

180. Laruelle M. Imaging synaptic neurotransmission with *in vivo* binding competition techniques: A critical review. J Cerebral Blood Flow Metab 2000;20:423–451.

181. Benfeldt E, Serup J, Menne T. Microdialysis vs. suction blister technique for *in vivo* sampling of pharmacokinetics in the human dermis. Acta Dermato-Venereolog 1999;79: 338–342.

182. Lefkowitz DL, Lefkowitz SS. (2001) Macrophage-neutrophil interaction: A paradigm for chronic inflammation revisited. Immunol Cell Biol 2001;79:502–506.

183. Maurer MH, et al. The proteome of human brain microdialysate. Proteome Sci 2003;1:7–15.

184. Sjögren F, Svensson C, Anderson C. Technical prerequisites for *in vivo* microdialysis determination of interleukin-6 in human dermis. Br J Dermatol 2002;146:375–382.

185. Ao X, Rotundo RF, Loegering DJ, Stenken JA. *In vivo* microdialysis sampling of cytokines produced in mice given bacterial lipopolysaccharide. J Microbiol Methods 2005;62:327–336.

186. Sopasakis VR, et al. High local concentrations and effects on differentiation implicate interleukin-6 as a paracrine regulator. Obesity Res 2004;12:454–460.

187. Winter CD, et al. A microdialysis method for the recovery of IL-1beta, IL-6 and nerve growth factor from human brain *in vivo*. J Neurosc Methods 2002;119:45–50.

188. Rosendal L, et al. Increase in interstitial interleukin-6 of human skeletal muscle with repetitive low-force exercise. J Appl Physiol 2005;98:477–481.

189. Heinemeier K, Langberg H, Olesen JL, Kjaer M. Role of TGF-beta1 in relation to exercise-induced type I collagen synthesis in human tendinous tissue. J Appl Physiol 2003;95:2390–2397.

190. Langberg H, Olesen JL, Gemmer C, Kjaer M. Substantial elevation of interleukin-6 concentration in peritendinous tissue, in contrast to muscle, following prolonged exercise in humans. J Physiol 2002;542:985–990.

191. Lin Y-C, Hesketh PJ, Lunte SM, Wilson GS. A micromachined diaphragm micropump. Proc—Electrochem Soci 1995; 95–27. 67–72.

192. Hesketh PJ, et al. Biosensors and microfluidic systems. Tribology Issues and Opportunities in MEMS, Proceedings of the NSF/AFOSR/ASME Workshop on Tribiology Issues and Opportunities in MEMS, Columbus, Ohio, Nov. 9–11, 1997, 1998; pp 85–94.

193. Zahn JD, Trebotich D, Liepmann D. Microfabricated microdialysis microneedles for continuous medical monitoring. Proceedings of the 1st Annual International IEEE/EMBS Special Topics Conference on Microtechnologies in Medicine & Biology. October 12–14, 2000, Lyon, France; 2000 pp 375–380.

194. Talbot D, Liepmann D, Pisano AP. Microfabricated polysilicon microneedles for minimally invasive biomedical devices. Biomed Microdevices 2000;2:295–303.

195. Bergveld P, et al. Microdialysis based lab-on-a chip, applying a generic MEMS technology. Comprehen Analy Chem 2003;39:625–663.

196. Carneheim C, Ståhle L. Microdialysis of lipophilic compounds: a methodological study. Pharmacol Toxicol 1991; 69:378–380.

197. Mueller M, et al. *In vivo* characterization of transdermal drug transport by microdialysis. J Controlled Release 1995;37: 49–57.

198. Trickler WJ, Miller DW. Use of osmotic agents in microdialysis studies to improve the recovery of macromolecules. J Pharmaceut Sci 2003;92:1419–1427.

199. Kurosaki Y, Nakamura S, Shiojiri Y, Kawasaki H. Lipomicrodialysis: a new microdialysis method for studying the pharmacokinetics of lipophilic substances. Biolog Pharmaceut Bull. 1998;21:194–196.

200. Khramov AN, Stenken JA. Enhanced microdialysis extraction efficiency of ibuprofen in vitro by facilitated transport with β-cyclodextrin. Analy Chem 1999;71:1257–1264.

201. Kjellstrom S, et al. Online coupling of microdialysis sampling with liquid chromatography for the determination of peptide and non-peptide leukotrienes. J Chromatog A 1998;823:489–496.

202. Rekharsky MV, Inoue Y. Complexation thermodynamics of cyclodextrins. Chem Rev 1998;98:1875–1917.

203. Stenken JA, Chen R, Yuan X. Influence of geometry and equilibrium chemistry on relative recovery during enhanced microdialysis. Anal Chim Acta 2001;436:21–29.

204. Khramov AN, Stenken JA. Enhanced microdialysis recovery of some tricyclic antidepressants and structurally related drugs by cyclodextrin-mediated transport. Analyst 1999; 124:1027–1033.

205. Ward KW, et al. Enhancement of in vitro and *in vivo* microdialysis recovery of SB-265123 using intralipid and encapsin as perfusates. Biopharmaceut Drug Disposition 2003;24:17–25.

206. Pettersson A, Markides K, Bergquist J. Enhanced microdialysis of neuropeptides. Acta Biochim Polon 2001;48: 1117–1120.

207. Pettersson A, et al. A feasibility study of solid supported enhanced microdialysis. Analy Chem 2004;76:1678–1682.
208. Ao X, Sellati TJ, Stenken JA. Enhanced microdialysis relative recovery of inflammatory cytokines using antibody-coated microspheres analyzed by flow cytometry. Anal Chem 2004;76:3777–3784.
209. Kissinger CB, Kissinger PT. Can preclinical ADMET-PK now be done more efficiently and effectively? Preclinica 2004;2: 319–323.
210. Olson DL, Lacey ME, Sweedler JV. High-resolution microcoil NMR for analysis of mass-limited, nanoliter samples. Anal Chem 1998;70:645–650.
211. Wolters AM, Jayawickrama DA, Sweedler JV. Microscale NMR. Curr Opin Chem Biol 2002;6:711–716.
212. Khandelwal P, et al. Studying rat brain Neurochem using nanoprobe NMR spectroscopy: A metabonomics approach. Analy Chem 2004;76:4123–4127.

Further Reading

The most comprehensive sources for microdialysis sampling are the book and the two separate journal issues shown below.

Robinson T, Justice JB, editors. Microdialysis in the Neurosciences. Amsterdam (The Nethernand): Elsevier; 1991.
Lunte CE *Anal Chim Acta* 1999;379:227–369.
Elmquist WF, Sawchuk RJ. Microdialysis sampling in drug delivery. *Adv Drug Del Res* 2000;45:123–307.

See also ELECTROPHORESIS; GLUCOSE SENSORS; HYDROCEPHALUS, TOOLS FOR DIAGNOSIS AND TREATMENT OF; PHARMACOKINETICS AND PHARMACODYNAMICS.

MICROFLUIDICS

GLENN M. WALKER
North Carolina State University
Raleigh, North Carolina

INTRODUCTION

Microfluidics is the study and application of fluids at the microscale. The most common definition of the microscale is that one or more device dimension be in the range of 1–1000 μm. For reference, the diameter of an average human head hair is ~150 μm, the average thickness of a human fingernail is 360 μm, and the diameter of a human red blood cell is ~7 μm. Miniaturization technology, originally developed by the microelectronics industry, has been used to create microscale fluid components and complete microfluidic systems with pumps, valves, and filters, incorporated onto single microchips have been demonstrated.

By applying the analogy of the microelectronics industry (i.e., continuously incorporating more features into smaller areas) a logical application of microfluidics is to create lab-on-a-chip (LOC) systems. Lab-on-a-chip systems, also known as micro-total-analysis systems (μTAS), incorporate the functionality of biology or chemistry laboratories onto a single microfabricated chip. Ideally, a LOC system would be able to execute all of the tasks routinely performed in a biology or chemistry laboratory, such as sample preconditioning, mixing, reaction, separation, and analysis. Labor- and time-intensive procedures would be reduced to instant results derived from a series of automated steps performed on a LOC.

Microscale fluid handling confers many advantages over traditional lab operations (1). First, fluid quantities ranging from picoliters to microliters are used, thus reducing the amount of sample required for tests. Second, the amount of time required to perform some analyses (e.g., capillary electrophoresis) is reduced to seconds, which means analyses can be conducted many times faster than with traditional methods. Third, devices can be manufactured using microfabrication technology, which translates into reduced cost per device; disposable LOC systems can easily be envisioned.

In general, microfluidic devices are in early stages of development and are most often found in academic research laboratories. However, the benefits of these systems have been exploited to develop new medical devices for clinical diagnostics and point-of-care testing. Commercial examples of devices that make use of LOC concepts are discussed at the end of this article.

THEORY

Fluid Mechanics

The term microfluidics encompasses both liquid and gas behavior at the microscale, even though in most applications the working fluid is a liquid. All of the concepts discussed here are directed toward liquids. Other works are available which provide information on gas behavior at the microscale (2).

Fluid behavior at the microscale is different from that commonly observed in everyday experiences at the macroscale, owing primarily to the very low Reynolds (Re) numbers of the flow regime plus the large surface area/volume (SAV) ratios of the flow domain. As a consequence, viscous forces and surface tension effects become dominant over fluid inertia, and transport phenomena are purely diffusive.

Fluid flow at the microscale is typically laminar. Fluid flows are classified based on their flow regime, which can be predicted with the Re number. The Re number is the ratio of inertial forces to viscous forces and can be calculated with the equation

$$\mathrm{Re} = \frac{\rho V D_h}{\mu} \tag{1}$$

where ρ is the fluid density, V is the characteristic fluid velocity, D_h is the hydraulic diameter of the microchannel, and μ is the fluid viscosity. Fully developed fluid flow in a channel of circular cross-section is considered laminar if the Re number is <2100. For Re numbers between 2100 and 2300, the flow is considered transitional: it shows signs of both laminar and turbulent flow. A Re number >2300 indicates turbulent flow.

Laminar flow is predictable in the sense that the trajectories of microscopic particles suspended in it can be accurately predicted (Fig. 1a). Particles suspended in a turbulent fluid flow behave chaotically and their position as a function of time cannot be accurately predicted

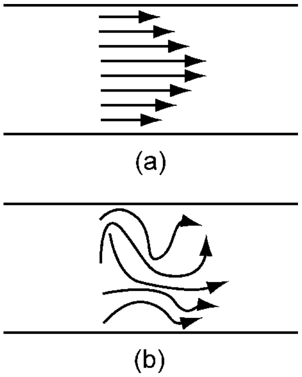


Figure 1. (a) Particles suspended in a laminar flow within a straight microchannel follow straight trajectories. (b) Particles suspended in a turbulent flow within a straight microchannel do not follow straight trajectories unless they are very close to the wall.

(Fig. 1b). Fluid flow that has a Re number <1 is also known as viscous flow, creeping flow, boundary-layer flow, or Stokes flow.

Low Re number flow is best visualized by imagining how honey (or any viscous substance) behaves when poured or stirred. For example, water flowing in microchannels will generally have a Re number <1. In this case, water will behave like a very viscous liquid (i.e., like honey). An important point to make here is that the properties of water do not change at the microscale; rather the microscale dimensions involved make the water *appear* more viscous than what we are accustomed to at the macroscale. An excellent description of low Re number environments has been given by Purcell (3). Very viscous fluid flows have certain characteristics: the flow is reversible, mixing is difficult, and flow separation does not occur (4).

Reversibility is the ability of a suspended particle in a fluid to retrace its path if the flow is reversed. This is a result of the minimal inertia (i.e., low Re number) present in fluid flows at the microscale. Figure 2 shows the path a suspended microscopic particle might take in forward and reverse flow.

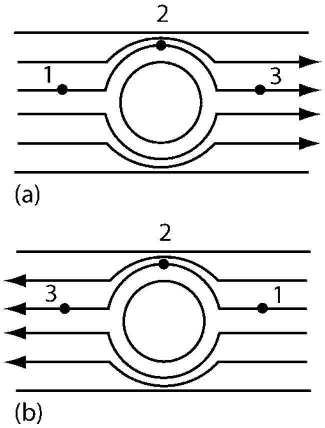


Figure 2. (a) A suspended particle in laminar flow around an obstacle in a microchannel. (b) If the flow is reversed, the particle will retrace the same path.

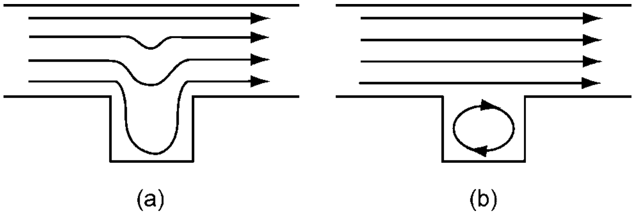


Figure 3. (a) At low velocities (low Re numbers), flow separation will not occur in microchannels. (b) At higher velocities (larger Re numbers), flow separation may become apparent.

A second characteristic of microscale fluid flow is a lack of flow separation. Flow separation is commonly observed in the form of vortices, which are recirculating flows separate from the main flow. Because of the low Re number environment, vortices usually will not form within microfluidic channels, as shown in Fig. 3a. Separation will only occur in flows wherein inertial forces are significant relative to viscous forces (Re > 1). Figure 3b is a qualitative sketch of flow separation in a cavity.

The third characteristic of microscale fluid flows is inefficient mixing as a result of very low *Re* number flow, and thus negligible inertia. Low inertia means that stirring is not effective and that mixing must be accomplished by diffusion. At the macroscale, stirring minimizes the diffusion distances between two or more liquids by distributing "folds" of the liquids throughout the volume. Microscale methods of mixing have been developed that take advantage of the unique properties of the scale and improve the efficiency of mixing over simple diffusion; examples include using three-dimensional (3D) channel geometries, patterned channel surfaces, and pulsatile flow (5).

Figure 4 shows two streams flowing down a microchannel side-by-side. Because of the low Re number environment the streams will only mix by diffusion. If the flowrate is slow enough, the streams will eventually become uniformly mixed across the whole microchannel width.

The hydraulic diameter, D_h, of a microchannel is determined by its cross-sectional geometry and can be calculated with the equation

$$D_h = \frac{4A}{P} \tag{2}$$

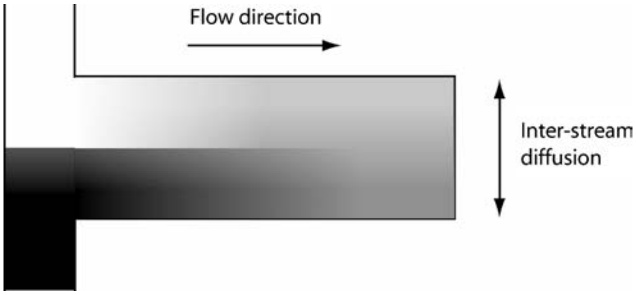


Figure 4. Two streams flowing in a microchannel will only mix by diffusion. Note that the concentration across the width of each half of the main microchannel is not constant as diffusional mixing progresses.

Table 1. Diffusion Coefficients for Biologically Important Molecules in Water[a]

Molecule	T, °C	D, $\mathrm{cm^2 \cdot s^{-1}}$	Diffusion time, s
Cl^-	25	2.03×10^{-5}	0.02
O_2	18	2×10^{-5}	0.03
K^+	25	1.96×10^{-5}	0.03
Na^+	25	1.33×10^{-5}	0.04
Glucose	20	6×10^{-6}	0.08
Lactose	20	4.3×10^{-6}	0.12
Insulin	20	1.5×10^{-6}	0.33
Hemoglobin	20	6.3×10^{-7}	0.79
Urease	20	3.4×10^{-7}	1.47

[a] All values are from Ref. 6. The time for each particle to diffuse 10 μm is shown for comparison.

where A and P are the microchannel cross-sectional area and wetted perimeter, respectively. The hydraulic diameter is often used to calculate important flow characteristics for noncircular microchannel cross-sections.

At microscale dimensions diffusion is an effective mechanism for transporting molecules because of the relatively short distances involved. Particles diffuse from areas of high concentration to areas of low concentration and will eventually diffuse to uniform concentration throughout a given volume. The mean distance, d, a particle travels in a time, t, can be predicted with the equation

$$d^2 = 2Dt \tag{3}$$

where D is the diffusion coefficient of the particle. Diffusion times are proportional to the square of distance, which means that particles can diffuse across microscale distances within a particular medium in a matter of seconds. Table 1 lists representative molecules of biological significance and their diffusion coefficients.

The SAV ratios become very large at the microscale. Typical SAV ratios for macroscale containers such as Petri dishes or culture flasks are ~10 $\mathrm{cm^{-1}}$, while they are ~800 $\mathrm{cm^{-1}}$ for microfluidic channels. Increased SAV ratios allow diffusion-limited processes, such as immunoassays to become much more efficient at the microscale because of the increased surface area available for binding. Large SAV ratios also allow rapid heat radiation from microscale fluid volumes and efficient gas exchange with the ambient atmosphere and fluid in microchannels (assuming the microchannel is made of a gas-permeable material). Enhanced gas transport is a critical ingredient for cell culture in microscale environments. One drawback of large SAV ratios is that evaporation becomes a significant problem.

The surface tension of a liquid becomes increasingly important at very small dimensions. To visualize this, think of a liquid surface as an elastic skin. If a slit were made in that skin, a certain amount of force per unit length would be required to hold the two sides of the slit together. The amount of force required to hold the two sides together is called the surface tension. Because the liquid surface is under tension, liquid confined by the surface (e.g., a raindrop) will experience an internal pressure. This pressure is called the Young–LaPlace pressure. Smaller fluid volumes result in larger SAV ratios, thus increasing the internal pressure. The pressure within a drop of liquid can be calculated with the formula

$$\Delta P = \frac{2\gamma}{R} \tag{4}$$

where γ is the liquid surface energy and R is the radius of the drop. At microscale dimensions, significant pressures can be created by surface tension. A common result of the pressure difference of an air/liquid interface is the capillary effect: a pressure difference across the interface propels liquid through a small diameter capillary or microchannel.

The capillary effect also depends on the contact angle of the microchannel surface. Hydrophobic surfaces (e.g., polymers) have contact angles >90° and hydrophilic surfaces (e.g., glass) have contact angles <90°. Microfluidic devices with hydrophilic surfaces can be filled via capillary action. The pressure difference at an air–liquid interface within a microchannel with square cross-sectional area can be calculated with the formula

$$\Delta P = 2\gamma\left(\frac{\cos(\theta_c)}{W} + \frac{\cos(\theta_c)}{H}\right) \tag{5}$$

where W and H are the microchannel width and height, respectively, and θ_c is the contact angle of the liquid on the internal microchannel walls. Conversely, equation 5 gives the pressure required to force water into a hydrophobic microchannel of rectangular cross-section.

Microfluidic Modeling

Microscale fluid flow can be modeled from either a macroscopic or microscopic vantage point. Macroscopic modeling treats the fluid as a well-mixed volume while the microscopic view looks at how particles suspended in the fluid would behave under different flow conditions.

Macroscopic modeling, also called lumped modeling, uses conservation of mass to predict microfluidic system behavior. A pressure drop, ΔP, applied across a microchannel (or other conduit) with fluidic resistance Z, will induce a volumetric flow rate Q:

$$\Delta P = QZ \tag{6}$$

All microchannels have a fluidic resistance associated with them that depends on the geometry of the microchannel and the viscosity of the fluid. The fluidic resistance of a microchannel with a circular cross-section is given by

$$Z = \frac{8\mu L}{\pi R^4} \tag{7}$$

where μ is the fluid viscosity, L is the microchannel length, and R is the microchannel radius. The fluidic resistance of a microchannel with a rectangular cross-section is given by

$$Z = \frac{4\mu L}{ab^3} f\left(\frac{a}{b}\right)^{-1} \tag{8}$$

where $f(a/b)$ is calculated with the formula

$$f\left(\frac{a}{b}\right) = \frac{16}{3} - \frac{1024b}{a\pi^5}\sum_{n=0}^{\infty}\frac{\tanh ma}{(2n+1)^5} \tag{9}$$

When calculating the resistance of microchannels with rectangular cross-section, μ is the fluid viscosity, L is the microchannel length, $2a$ and $2b$ are the microchannel width and height, respectively, and m is calculated with

$$m = \frac{\pi(2n+1)}{2b} \tag{10}$$

If the aspect ratio of the microchannel is very small (i.e., $2b \ll 2a$) then the simplified formula

$$Z = \frac{3\mu L}{4ab^3} \tag{11}$$

can be used. The general rule of thumb is that equation 11 should be used for microchannels with b/a <0.1. The resistance of other geometries can be found elsewhere (7).

In predicting microfluidic system behavior, the analogies to Kirchhoff's laws are used. The sum of pressure drops in a fluidic loop must be equal to zero; the total volumetric flowrate entering a node must be equal to the total volumetric flowrate leaving a node.

In contrast to the macroscopic view, microscopic modeling allows fluid behavior to be predicted. Specifically, the microscopic view allows the velocity profiles of a fluid flow to be calculated. Velocity profiles are plots that show the relative velocities of different portions of a fluid within a microchannel. Figure 1a is an example of a velocity profile.

The velocity of flow in a microchannel with circular cross-section varies radially and can be predicted with the formula

$$v(r) = \frac{R^2\Delta P}{4\mu L}\left(1 - \frac{r^2}{R^2}\right) \tag{12}$$

where μ is the fluid viscosity, L is the microchannel length, ΔP is the pressure drop, and R is the microchannel radius. The velocity profile of flow in a microchannel with rectangular cross-section varies along the height and width axes and can be predicted with the formula

$$v(x,y) = \frac{\Delta P}{2\mu L}\left(b^2 - y^2 - \frac{4}{b}\sum_{n=0}^{\infty}(-1)^n\frac{1}{m^3}\frac{\cos my \cosh mx}{\cosh ma}\right) \tag{13}$$

where μ is the fluid viscosity, L is the microchannel length, ΔP is the pressure drop, m is calculated from equation 10, and $2a$ and $2b$ are the microchannel width and height, respectively.

Microscopic modeling is performed when precise modeling of fluid behavior is needed. For example, cells attached to the wall of a microchannel might affect flow; modeling at the microscopic level would reveal any perturbations of the flow caused by the cell. In contrast, macroscopic modeling is performed when the behavior of the entire microfluidic system is needed. For example, fluid flow in many parallel microchannels might be required. Macroscopic modeling would reveal the relative flowrates through each microchannel and provide the microchannel dimensions needed to guarantee equal flow through each.

PUMPING FLUIDS

Fluids are pumped through microfluidic channels by creating gradients; the two most common types being pressure and electrical. Other types of gradients and their applications are discussed elsewhere (8).

Pressure gradients are the most common method used to pump fluid. Pressure is applied to one end of a microchannel which causes the fluid to flow down the pressure gradient. Common methods for creating a pressure gradient include pumps or gravity. Most methods for creating pressure-driven flow use macroscale pumps attached to the microfluidic device via tubing. Ideally, pumps should be incorporated on-chip to realize the ultimate vision for LOC devices. Many types of microfluidic pumps have been demonstrated and they presently constitute an active area of research (9).

Pressure-driven flow is attractive for use in microfluidics because it is easy to set up and model. Some drawbacks for using pressure-driven flow are sensitivity to bubbles, sensitivity to motion (via the tubing connecting pumps to the microfluidic device), and parabolic flow profiles. Shear stress is proportional to the pressure drop across a microchannel, which should be taken into account when manipulating cells.

The other common way to pump fluids is to use electrical gradients. This method of pumping fluid is only practical at the microscale level because of the large electric fields and SAV ratios required. Pumping via electric fields is called electrokinetic flow and is based on two phenomena: electrophoresis and electroosmosis. Electrophoresis operates on the principle that charged particles in an electric field will feel a force proportional to the field strength and their charge. The particles will move through the electric field toward the pole of opposite charge. Larger particles move slower than smaller particles because of the drag produced by moving through a fluid. Figure 5a shows an example of electrophoretic flow.

A charged particle in an electric field of strength E will travel with a velocity equal to

$$v = m_{\mathrm{ep}}E \tag{14}$$

where m_{ep} is the electrophoretic mobility. Electrophoretic velocities are typically much smaller than the velocities caused by electroosmosis.

Electroosmosis will only function in the presence of an electric double layer at the surface of the microchannel.

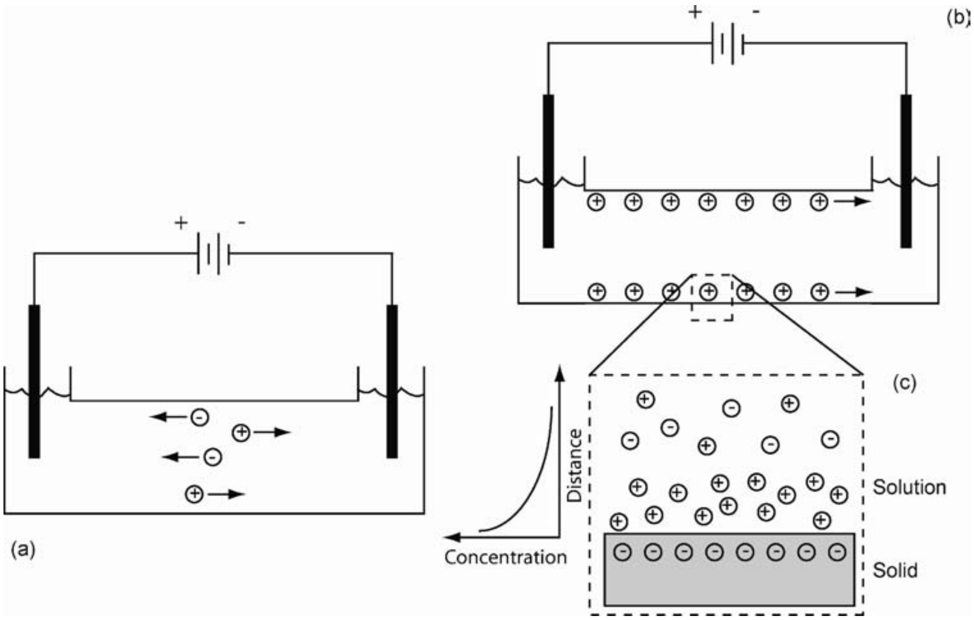


Figure 5. (a) Electrophoresis. Charged particles will move toward oppositely charged poles in an elec- tric field. (b) Electroosmosis. Charges lining a microchannel sur- face will move with an applied electric field, thus inducing bulk flow via momentum transfer within the liquid. (c) An electric double layer forms at charged microchannel surfaces; the layer thickness is called the Debye length.

An electric double layer forms at a charged surface when oppositely charged particles from an electrically neutral liquid gather at the surface. The thickness of the electric double layer is known as the Debye length; the concentration of charged particles at a surface falls off rapidly as a function of distance and is shown in Fig. 5c. When an electric field is applied across the length of the microchannel, the ions gathered at the microchannel surface begin to slide toward the oppositely charged pole, as shown in Fig. 5b. As the ions slide, they drag their neighbors within the bulk liquid, toward the middle of the microchannel. The friction between subsequent sliding layers of ions causes the bulk fluid to begin moving.

If the Debye length is much less than the characteristic dimensions of the microchannel, then the bulk fluid velocity can be predicted with the equation

$$v = m_{\mathrm{eo}}E \tag{15}$$

where E is the electric field strength and m_{eo} is calculated with

$$m_{\mathrm{eo}} = \frac{\varepsilon\zeta}{4\pi\mu} \tag{16}$$

where ε is the dielectric constant of the fluid, ζ is the zeta potential of the surface, and μ is the fluid viscosity.

Electrokinetic flow is attractive because it only requires the integration of electrodes in a microfluidic device, which is straightforward by microfabrication standards. Electrokinetic flow is therefore amenable to interfacing with electronic control circuitry. Electrokinetic flow also results in a blunt flow profile, which reduces the distortion of transported samples. Lastly, electrokinetic flow has very rapid response times, since the electrodes are integrated on chip, and are generally insensitive to movement off chip. Drawbacks to electrokinetic flow include fouling of the electrodes, which reduces electric field strength and therefore flowrate. Also, protein adsorption to microchannel surfaces affects the Debye layer, and in turn flow. Unintended side effects from electric fields on biological cells within the microfluidic device may also exist. Figure 6 shows the difference between the parabolic flow profile of pressure-driven flow and the blunt profile of electrokinetic flow.

Other methods of fluid flow based on surface tension, heat, and evaporation, have also been demonstrated. However, these methods have yet to be widely adopted and it is

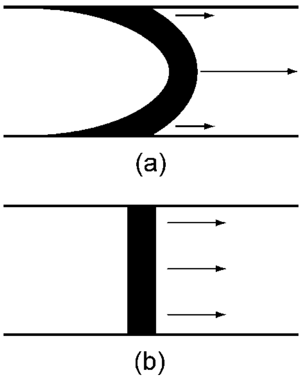


Figure 6. (a) Pressure-driven flow is parabolic; the middle of the stream moves faster than regions near the wall. (b) Electrokinetic flow is blunt; all parts of the stream move at equal velocity. Note that residual pressures can cause the profile to become slightly parabolic in the direction of the negative pressure gradient.

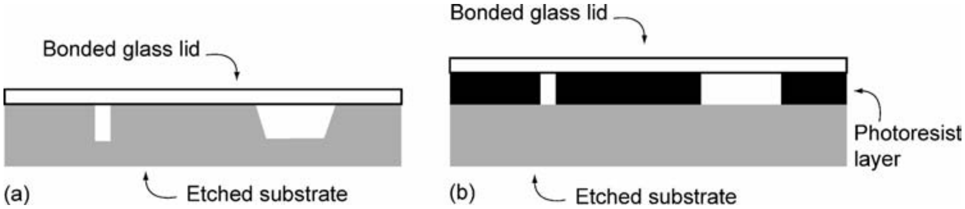


Figure 7. (a) Material is etched from a substrate and an enclosed structure is formed by bonding the etched substrate to a glass lid. (b) Walls of microchannels are built on top of a substrate and a glass lid is placed on top of the photoresist to form an enclosed structure.

unclear if they will prove more attractive than pressure-driven or electrokinetic flow.

FABRICATION

Fabricating microfluidic channels in traditional microfabrication materials, such as silicon and glass, can be achieved in two different ways. In the first, material is selectively removed, or etched, from a bulk substrate. The etched substrate is then bonded to another material (e.g., glass or silicon), which may have access holes or other features embedded in it. The result is an enclosed channel structure, as shown in Fig. 7a. The second method is to selectively add material to a substrate, and then bond another substrate to it. This method will also form enclosed channel structures as shown in Fig. 7b.

Photolithography is a fundamental part of all microfabrication (Fig. 8). Light is used to project patterns onto a photosensitive chemical, called a photoresist. The photoresist can be either positive or negative. Light chemically alters positive photoresist and makes it soluble in a developer. Negative photoresist is cross-linked by light, which makes it insoluble in developer. The patterned photoresist can be used as an etch mask for substrates, producing microchannels like those shown in Fig. 7a. Patterned photoresist can also be used in subsequent steps to direct the patterning of other materials that cannot be directly patterned with light. In-depth treatments of microfabrication techniques can be found elsewhere (10,11).

Polymers have recently become popular alternatives to traditional (e.g., silicon or glass) microfabrication materials. Polymers can be used to make microchannels by using the same methods mentioned previously for silicon and glass. Polymers also have the advantage that they can be molded that makes them a cheaper alternative (relative to silicon or glass) for mass production.

Polymer microfluidic devices can be created by molding, hot embossing, injection molding, photopolymerization, and laser ablation or laser cutting. An attractive method for fabricating polymer microfluidic devices is to use a process known as micromolding (12). In this process, photolithography is used to make a pattern of the microchannels (called a master). The photoresist provides a positive relief from which a polymer mold can be cast. The polymer is poured over the master and allowed to cure. The polymer mold is then peeled from the master and either placed on a substrate or incorporated into a multilayer device. The two advantages of this microfabrication method are (1) no special microfabrication equipment is required, and (2) many inexpensive copies of a microfluidic device can be rapidly manufactured.

Drawbacks to using polymers include leaching of material into microfluidic channels, solvent incompatibility,

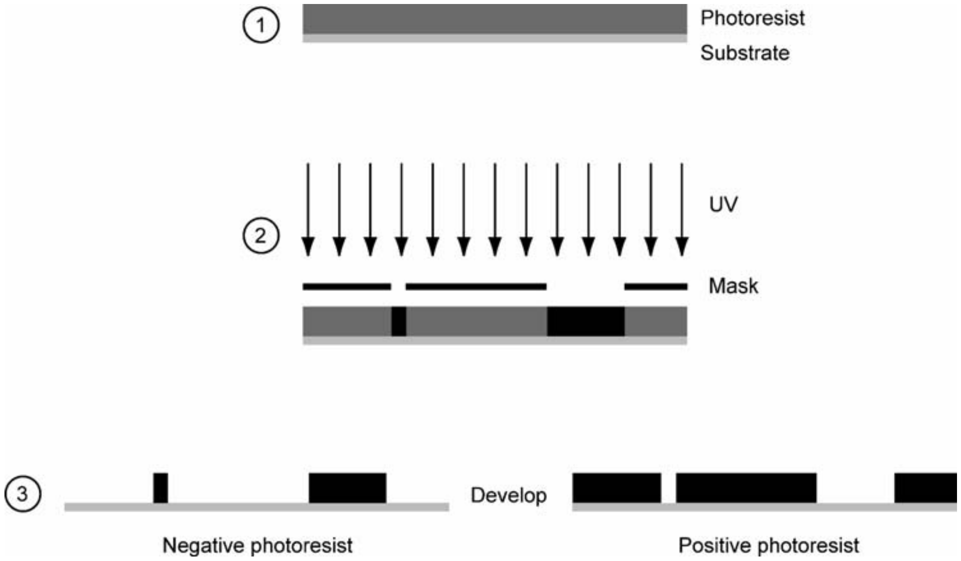


Figure 8. Photolithography requires an ultraviolet (UV) light source, a mask, and a photosensitive layer of material (i.e., photoresist). The photoresist is patterned with the UV light via a mask.

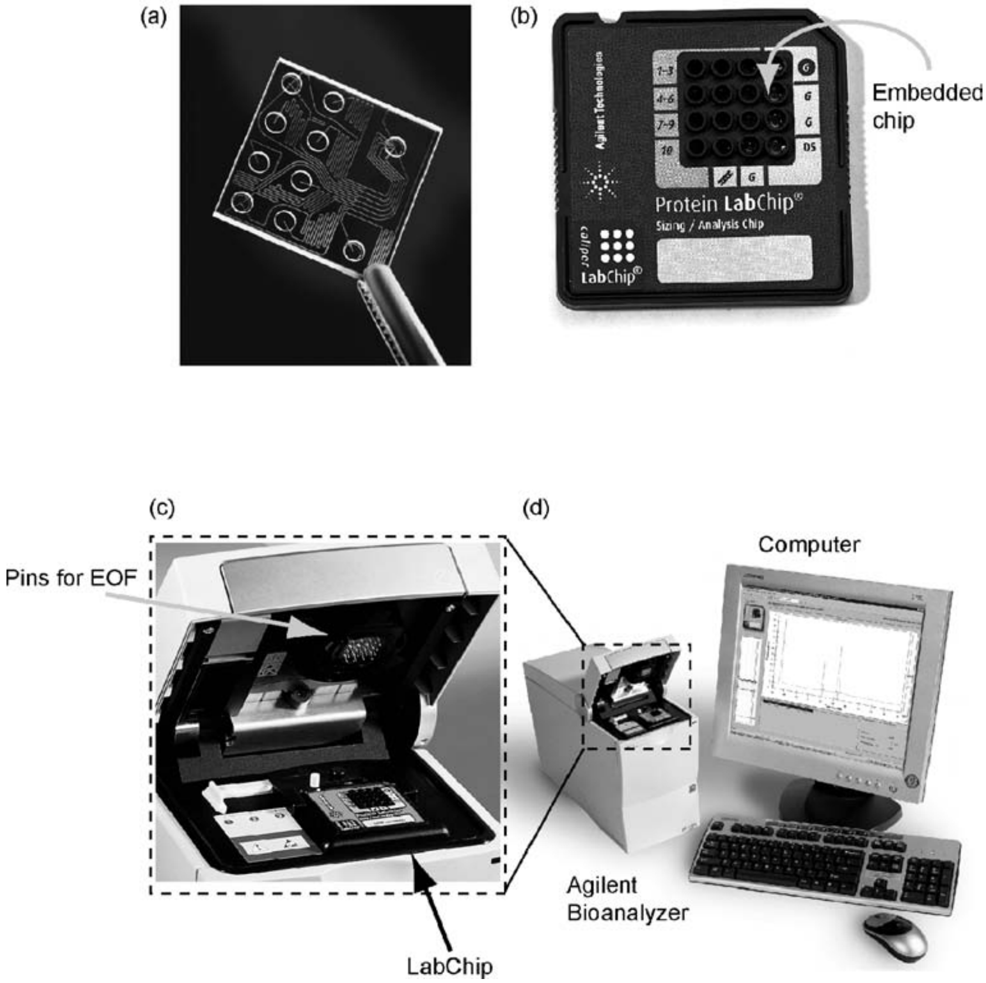


Figure 9. (a) The Bioanalyzer uses microfluidic chips etched from glass. (Images courtesy of Agilent Technologies). (b) The etched glass chips are encased in a plastic assembly that facilitates handling and limits contamination. Images courtesy of Agilent Technologies. (c) Samples are loaded onto the chip and the chip is loaded into the Bioanalyzer. (Images courtesy of Agilent Technologies). (d) The Bioanalyzer then performs an analysis on the samples. (Images courtesy of Agilent Technologies.)

and the ability of some substances to diffuse into the polymer. Also, surface treatments are occasionally required to make polymers compatible with electrokinetic flow; silicon or glass have an inherent surface charge that allows them to be used in electrokinetic flow applications.

BIOMEDICAL APPLICATIONS OF MICROFLUIDICS

Microfluidic concepts have already been incorporated in a variety of biomedical devices (13). One example of a microfluidic device now found in many biomedical labs is Agilent's Bioanalyzer. The Bioanalyzer system uses disposable chips etched in glass. The samples to be separated and reagents for the separation are loaded onto the chip via reservoirs. The chip is then placed in a reader, where electrokinetic flow is used to manipulate the samples in the reservoirs (Fig. 9).

Microfluidic capillary electrophoresis systems are becoming commonplace in laboratories. By using very small volumes for separation, joule heating from electrophoresis is rapidly radiated away from the gel. Efficient heat radiation allows larger voltages to be used which translates into faster separations. The shorter separation distances used also contribute to reduced analysis times. Microfluidic capillary electrophoresis systems allow DNA to be rapidly analyzed, and have highly reproducible results since the entire process is automated. Lastly, contamination is minimized because the devices are disposable.

Because of their small size, another attractive aspect of capillary electrophoresis (CE) systems is that they can be incorporated into LOC devices and made part of a complete system. An example that has been demonstrated in several research labs is a system that takes cells as inputs, lyses them, performs all necessary preprocessing, DNA amplification, and so on, and then performs the DNA separations, all with no human intervention (14).

Microfluidics are also being used in clinical devices; devices for hematology and disposable assays for point-of-care diagnostics are among those now being researched and brought to market. A handheld point-of-care device made by i-STAT is an example of a clinical microfluidic device (Fig. 10). The handheld device quantifies analytes in

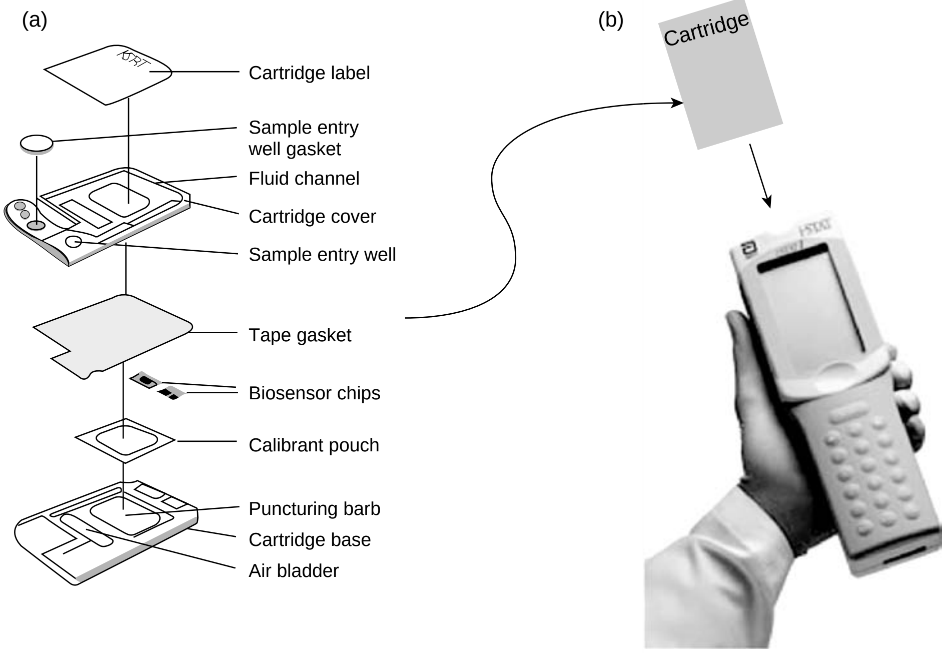


Figure 10. (a) The handheld point-of-care device manufactured by i-STAT performs analyses on blood samples contained in a disposable cartridge. Reprinted with permission from ACS (from Ref 15). Copyright 1998 American Chemical Sa. (b) Microfluidic cartridges are loaded with a sample and then plugged into the i-STAT handheld device. Image courtesy of Abbott Point-of-Care. The cartridge contains all necessary microfluidic control and sensing components that are then actuated by the handheld device.

blood samples that have been deposited on a disposable chip. The general procedure for operation is given below (15).

A patient's blood sample is deposited in a well on the disposable chip and then the well gasket is snapped shut (Fig. 10a). The microfluidic chip is then inserted into a handheld reader that performs an automated analysis of the blood sample, as shown in Fig. 10b. On-chip biosensors are automatically calibrated and checked for accuracy with an on-chip packet of calibrated solution. Once their accuracy has been determined, the calibration solution is flowed to the waste compartment. The blood sample is then flowed over the biosensors and the concentrations of different analytes are displayed on the handheld device screen within a few minutes. Diaphragm pumps are used to move the fluid. A variety of chips are available for different assays, including electrolytes and blood gases.

CONCLUSION

Microfluidics is the study and application of fluids at the microscale. Techniques used by the microelectronics industry have been adapted to facilitate the creation of micron-size channels capable of carrying fluid. The physical behavior of fluid at the microscale differs from behavior observed at the macroscale in everyday experience. The miniaturization of fluid handling has allowed LOC devices to be created in which all of the procedures of a traditional chemistry or biology lab are performed automatically in a single microfabricated chip. Lab-on-a-chip devices will allow new clinical and research tools to be developed.

BIBLIOGRAPHY

1. Brody JP, et al. Biotechnology at low Reynolds numbers. Biophys J 1996;71(6):3430–3441.
2. Karniadakis GE, Beskok A. Micro Flows. 2nd ed. New York: Springer-Verlag; 2001. p 360.
3. Purcell EM. Life at low Reynolds number. Am J Phys 1977;45(1):3–11.
4. Meldrum DR, Holl MR. Tech.Sight. Microfluidics. Microscale bioanalytical systems. Science 2002;297(5584):1197–1198.
5. Nguyen NT, Wu ZG. Micromixers—a review. J Micromech Microeng 2005;15(2):R1–R16.
6. Stein WD. Channels, Carriers, and Pumps: An Introduction to Membrane Transport. San Diego: Academic; 1990.
7. Shah RK, London AL. Laminar Flow Forced Convection in Ducts. New York: Academic; 1978.
8. Stone HA, Stroock AD, Ajdari A. Engineering flows in small devices: Microfluidics toward a lab-on-a-chip. Ann Rev Fluid Mech 2004;36:381–411.
9. Laser DJ, Santiago JG. A review of micropumps. J Micromech Microeng 2004;14(6):R35–R64.
10. Kovacs G. Micromachined Transducers Sourcebook. Boston: WCB McGraw-Hill; 1998.
11. Madou M. Fundamentals of Microfabrication. 2nd ed Boca Raton(FL): CRC Press; 2002.
12. McDonald J, et al. Fabrication of microfluidic systems in poly(dimethylsiloxane). Electrophoresis 2000;21(1):27–40.
13. Beebe DJ, Mensing GA, Walker GM. Physics and applications of microfluidics in biology. Ann Rev Biomed Eng 2002;4:261–286.
14. Waters L, et al. Microchip device for cell lysis, multiplex PCR, amplification, and electrophoretic sizing. Anal Chem 1998; 70(1):158–162.
15. Lauks IR. Microfabricated biosensors and microanalytical systems for blood analysis. Acc Chem Res 1998;31(5):317–324.

See also DRUG DELIVERY SYSTEMS; NANOPARTICLES

MICROPOWER FOR MEDICAL APPLICATIONS

Ji Yoon Kang
Korea Institute of Science and
Technology
Seoul, Korea

INTRODUCTION

Generally, micropower is the local generation of electricity by small-scale generators, which locates the end point. As the recent development of the microelectromechanical system (MEMS), as well as CMOS electronics technology, has been reducing the size and cost of biomedical devices, the research of micropower became important for implantable biomedical devices since they require internal self-sustained power sources.

As for biomedical devices, micropower is an internal or external power source to supply energy for active devices, which replaces an organ's function or treats diseases. The examples of active implantable devices that consume energy are cardiac pacemakers, cardiac defibrillators, muscle stimulators, neurological stimulators, cochlear implants, and drug pumps (1). Hence, in this article the term micropower describes rather tiny power supplying devices for miniaturized sensors, actuators, and electric devices, whose size is $>$ or $< 1\ \text{cm}^3$.

The low power electrical actuator, such as a pacemaker or neuronal stimulator, requires tens of microwatts intermittently and their power source is usually a lithium iodine battery that lasts from 5–8 years. Usually, the stand-by current of a pacemaker is $\sim 1\ \mu\text{A}$ in waiting mode and its pulse current is ~ 6 mA. One example of a specification for a pacemaker pulse is in the range of $25\ \mu\text{J}$ (~ 11 mA at 2.2 V with a 1 ms discharge) and the capacity of the battery is 2 Ah at typical rating (2). The volume of a pacemaker is ~ 20 mL and the volume occupied by the battery is about one-half of the total volume, 10 mL. Hence, the energy density (energy/volume) and reliability are important factors in the lifetime of the device.

An internal battery that is hermetically sealed in these devices can operate them with low power consumption; however, other implantable devices have radically different power requirements. Implantable cardioverter defibrillators demand the energy of 15–40 J providing six orders of magnitude larger than that of a pacemaker even though the pulses are less frequent. The current from a lithium silver vanadium battery is charged in an internal capacitor and the pulse of 1–2 A of current is fired. Electromechanical actuator like drug pumps demand more current than a lithium ion battery can deliver since it needs to overcome the high pressure in the chamber. The examples of drug pumps are insulin pumps, pain reliever, and an cerebrospinal fluid pump. A high current implantable battery should have low source impedance, such as lithium thionyl chloride, lithium carbon monofluoride, or lithium silver vanadium oxide.

Other future application are in wireless sensors, including physiological, chemical, or physical sensors embedded in an encapsulated environment. Miniaturized sensor consumes $< 100\ \mu\text{W}$ and a radio frequency (rf) transmitter consumed ~ 10 mW intermittently. Since most of the power is used for communication, some research groups are developing several low power wireless transmission protocols (3,4). Hence, less power will be necessary for a sensor transmitter as technology evolves.

Some groups investigated more efficient and reliable batteries. To enhance their efficiency and lifetime, potential alternatives of the conventional batteries studied, such as a microfabricated battery, microfabricated fuel cell, and biofuel cell. Microelectrical system technology reduces the size of the primary battery and microfluidic galvanic cell (5), water activated microbatteries (6), and Li-ion microbatteries were demonstrated. The fuel cell attracts much attention since it has a high efficiency, high power, and low pollution rate. Research on fuel cells focus on high power applications, such as the automobile and portable electronics, like laptop computers and cellular phones (7). Recently, micromachining technologies employed as a method to fabricate miniature fuel cells (8–11) and their size became smaller than a button cell battery (12) with high power. Enzyme-based glucose/O_2 biofuel cells were reported by several groups (13) and a miniaturized all (14) was reported that is $< 1\ \text{mm}^3$, with although a power of $4.3\ \mu\text{W}$. Since glucose is available in all tissues and organs, it is advantageous in implanted medical sensortransmitters.

Although a primary battery as well as a rechargeable battery is an important tool that supplies reliable power to implant devices, the continuous power of the battery decreases with time, and after 5 years they will not supply enough power (15). Power delivery with an rf transmission can extend the lifetime and continuously deliver high power. In the case of a cochlear implant, an external device provides power and data through electromagnetic field coupling; however, it needs accurate positioning of the external device and may cause rf interference and heating of the tissue.

Therefore, many research groups are paying attention to microfabricated power scavenging devices as an auxiliary power source to recharge the battery with no external power. Ambient energy sources are body heat or movement of the human body. Piezoelectric material (16–18), capacitance change (19–24), and inductive coil (25–27) convert vibration or human motion into electrical energy. The generation by high frequency vibration is not suitable for implant devices since vibration of the human body is in the range of tens of hertz. Hence, energy conversion using vibration of low frequency can be integrated with implant devices.

Thermoelectric generators that convert temperature differences of the human body or combustion engine to electricity were reported (28,29). Another conversion method is the thermophotovoltaic power generator (30,31), which combines the combustion engine and solar cell. However, integrating a power generation device into an implant device has some limitations due to biocompatibility. Power generation with a high temperature like the combustion engine or thermovoltaic metoid, cannot be implemented in the inside of the human body due to the heating of tissues. Thermoelectric generation using body heat is promising for the implantable device. However, this article includes the review on the other portable power sources like the micro-combustion engine, because those are also useful for a portable diagnosis system. The recent development of microfluidics and miniaturized biosensors enables point-of-care testing devices to be on the market in the near future. A microheat engine is highly efficient in energy conversion and will be useful as a portable medical equipment.

A good review article on micropower for wireless sensor networks was reported (20,32). It lists the candidates of portable power sources and compares the energy density for the battery and power density for a power generator. This article reviews the existing and potential micropower sources in view of medical applications from tiny sensors embedded in the human body to portable medical electronic devices.

MICROBATTERY

For a long time, the battery was a major energy source in portable electronic devices and it has evolved from Zn/MnO_2 to the Zn/air cell since 1900. Electrochemical power sources were developed in response to the needs of the flashlight, automotive starter, mobile electronics, and laptop computers. These days, there is a tremendous need for portable electronics demanding smaller, lighter, and longer lived batteries. Hence, many researches are on the way to making microfuel cell or microfabricated batteries using various kinds of electrochemical power. This section will briefly review microbatteries including fuel cells, biofuel cells, and micromachined batteries.

Microfuel Cell

Although the energy density of conventional batteries has been increasing from 500 $Wh \cdot L^{-1}$ for the Ni/Cd battery to 1500 $Wh \cdot L^{-1}$ for the Li/C–CoO_2 battery, there is a large jump for the air-cathode fuel cells of 4500 $Wh \cdot L^{-1}$ using hydrogen, hydrocarbon, and metals (7). The proton exchange membrane fuel cell (PEMFC) depicted in Fig. 1 is one of the promising techniques for fuel cell, which was implemented in miniaturization. Power sources for the automobile or the portable electronics of a huge market have been a main concern of fuel cell research because of the advantage of high efficiency and easy rechargeability. However, these days, in response to the demand of low power application, many miniaturized fuel cells are under study. Miniature fuel cells with a series path in a flipflop configuration was fabricated in a planar array and a four-cell prototype was produced, 40 $mW \cdot cm^{-1}$ (33). Yu (11) added microfluidic channels using anisotropic wet etching of silicon to the flipflop configuration and measured a peak power density of 190 $mW \cdot cm^{-2}$. They also reported that the flipflop fuel cell was constructed on printed circuit board (PCB) and that they achieved the area power density of >700 $mW \cdot cm^{-2}$. Wainright (12) fabricated on-board hydrogen storage with multiple coplanar fuel cells in series on ceramic substrates. They stored hydrogen in the form of the stabilized aqueous solutions of sodium borohydride ($NaBH_4$) or a metal hydride material, such as $LaAl_{0.3}Ni_{4.7}$. The energy density of microfabricated fuel cells did not exceed that of a Li/MnO_2 coin cell, but it had the advantage of higher power and compatibility with other microelectronic, microelectromechanical, or microfluidic devices. A polymer microfluidic channel was applied to a fuel cell using PDMS (10) and poly (methyl methacrylate) (PMMA) (8). They achieved a comparable area power density with a silicon based one. Whitesides and co-workers (9) also reported a membraneless vanadium redox fuel cell using a laminar flow property in a microfluidic channel with 200 $mW \cdot cm^{-2}$. As for the commercialization, MTI micro-fuel cells and Manhattan Scientifics are making portable fuel cells using direct methanol fuel cell (DMFC) and Medis technology announced direct liquid fuel cell (DLFC) for handheld devices. Miniaturized fuel cells are promising for implantable medical devices with a high power requirement and it has the advantage of a longer lifetime and less charging time than a conventional battery.

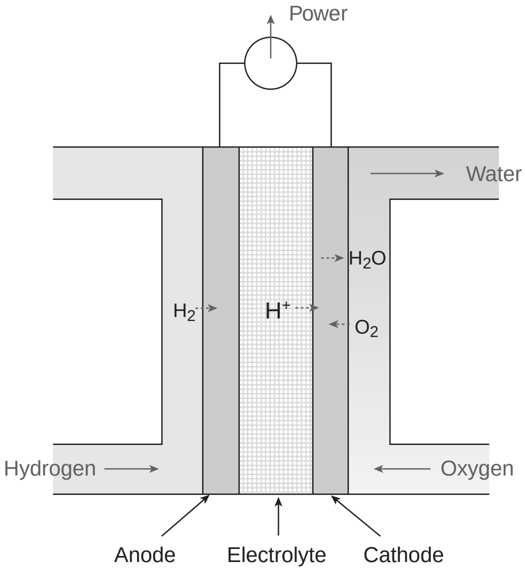


Figure 1. Schematic of a PEM fuel cell.

However, PEM fuel cells do not fit well with the implantable microdevices with high power and a long life application because the refill of a hydrogen fuel cell is not easy when it is sealed in the human body. As an alternative fuel cell, the biofuel cell is a strong candidate for a low power embedded device like a microbiosensor. Although a biofuel cell is not capable of high power, the easy availability of fuel (glucose) gives it a long operation time. Enzyme-based glucose/O_2 biofuel cells were studied since 1980s (13) and Heller and co-worker (34) demonstrated the feasibility of a membraneless miniature biofuel cell as an implanted micropower source. Its chemical reaction is described in equations 1 and 2.

$$\text{glucose} \rightarrow \text{glucolactone} + 2H^+ + 2e^- \quad (1)$$

$$O_2 + 4H^+ + 3e^- \rightarrow 2H_2O \quad (2)$$

Some fuel cell components, such as case, membrane, ion conductive electrolyte, and plumbing was removed in the biofuel cell and its size became <1 mm^2. The power of the biofuel is 4.3 μW with 0.52 V and it is suitable for an implanted devices because of tiny volume and abundant glucose inside of human body. Moor et al. (35) developed a microfluidic chip based ethanol–oxygen biofuel cell, which produced 18 μW with 0.34 V and is applicable to integrate the biofuel cell and the microfluidic chip.

Micromachined Battery

A microfabricated battery usually refers to a thin-film battery, however, recently some research groups are studying MEMS-based microbatteries. Integration of microbatteries with CMOS electronics or an MEMS device is easy and can be fabricated on a chip with a device. Since the microbatteries main concern is power output due to limitation of surface rather than the capacity, most research activities are focused on increasing power to out perform maintaining capacity.

As for thin-film batteries, Bates et al. (36) at Oak Ridge National Laboratory reported a thin-film secondary battery, which was made up of lithium and lithium ion. The thickness is tens of μm and the area is in the cm^2 range. Its continuous current output is $1\ mA \cdot cm^{-2}$. Pique et al. (37) constructed primary $Zn–Ag_2O$ and secondary Li ion microbatteries in plane using laser direct-write with a capacity of $100\ \mu Ah \cdot cm^{-2}$.

Other than classical thin-film batteries, several micromachined MEMS batteries were developed. They try to integrate power sources with microelectronic circuits and microsensors. Prof. Lin at UC Berkeley proposed a water-activated battery with 1.86 mWh in the area of 12 × 12 mm for lab-on-a-chip application (6) that overcomes the corrosiveness of the micromachined batteries with sulfuric acid and hydrogen peroxide (38). Andres (5) devised a pump integrated with a micropower source, in which microfluidic galvanic cells supplied power to heat up the two-phase fluid for pumping. The capacity of the micromachined battery is lower than that of a thin-film battery yet, its application is restricted to the integrated power for a micromachined implantable device.

MICROPOWER GENERATOR

A micropower generator scavenges energy from devices. The energy sources are mechanical (vibration and human body movement), thermal (temperature difference), and solar energy. Thermoelectric devices convert the temperature difference to electricity and the photovoltaic cell collects solar energy. A MEMS-based power generator scavenges energy from the vibration in the mechanical structure or human body movements.

Thermoelectric Generator

Thermoelectric generation using the Seebeck effect was widely studied. This effect was discovered in 1821 by the physicist, Jonn Seebeck. It is the same phenomenon with a thermocouple where the temperature difference produces electricity or work. Although a thermoelectric power generator is an old technique and is commercially available in various sizes in the market, recent research focuses on making miniaturized low power thermoelectric microgenerators using microfabrication. Cost-effective fabrication technology was developed using electroplated structures with an epoxy film (28) and nanowire arrays by electrochemical deposition was implemented to improve thermoelectrical properties (29). Reportedly, several companies announced a thermoelectric generator using body heat. Applied Digital Solutions (39) announced a thermoelectric generator called ThermoLife, which produced a power of 49 μW from a temperature difference of 5 °C in $0.5\,cm^2$. Biophan technologies (40) also announced a biothermal power source as shown in Fig. 2 for implantable devices like a pacemaker and defibrillator (41). They aim to produce 100 μW at 3 V with 1 °C temperature difference. A different scheme converting thermal energy to electricity is piezoelectric generator actuated by thermal expansion of two-phase working fluid (42). Although they produced up to 56 μW at its resonance frequency of 370 Hz, a practical application needs the careful design of a heat-transfer mechanism because the temperature is required to oscillate at a resonance frequency of structure. A thermoelectric generator has a well-established technology and its operation is relatively stable; however, note that the temperature difference inside the human body is $< 1\,°C$. The temperature gradient is maximum at the skin surface and will limit the location of the thermoelectric power generator.

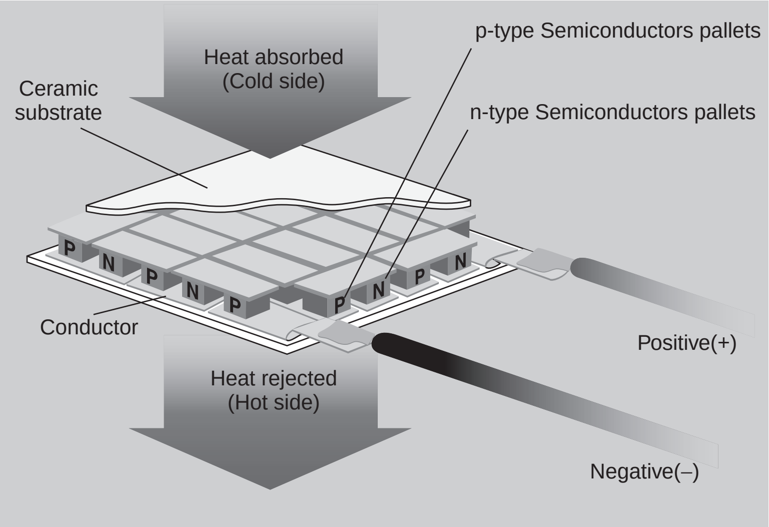


Figure 2. Principles of thermoelectric power source (Biophan).

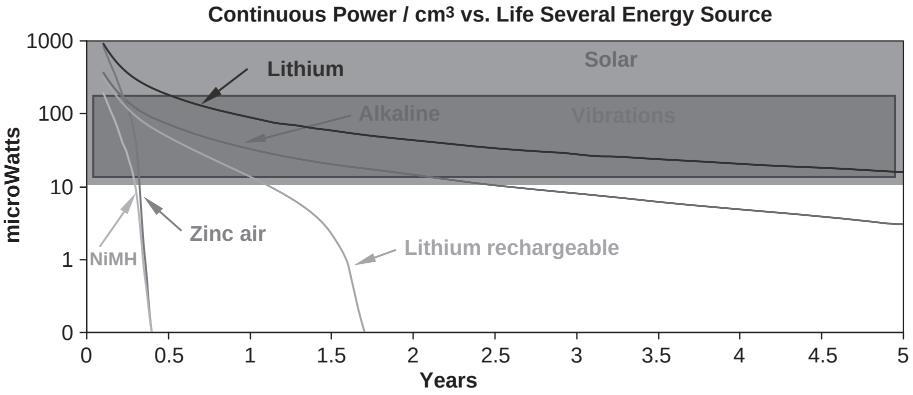


Figure 3. Comparison of power density from vibrations, solar and batteries (21).

Power Generation with Ambient Vibration

Conversion of a mechanical vibration to electric power has been studied from the mid 1990s (43) using MEMS, while thermal and solar energy were exploited to generate electricity long ago. They changed vibration to electrical energy using piezoelectric, electromagnetic, and electrostatic generations.

The frequency of vibration in an ambient environment ranges from 60 to 400 Hz and for the microwave oven the acceleration was 2.25 $\mathrm{m \cdot s^{-2}}$ at a resonance frequency of 120 Hz (15). Shad (21) suggested a graph comparing the power density of power scavenging and batteries as a function of time (Fig. 3). Power density in the vibration of machining center or microwave oven becomes larger than conventional batteries after 3 or 4 years, which means the waste vibration energy is not negligible. Williams and Yates (43) presented a general model in equation 3 for the power of external vibration as depicted in Fig. 4, when a mass is moved at a resonant frequency of ω_n.

$$P = \frac{\zeta_t m Y_o^2 \omega_n^3}{4(\zeta_t + \zeta_o)^2}\zeta \tag{3}$$

where m is mass, and Y_o is the maximum extent that the mass can move, ζ_t and ζ_o denote a damping factor both in the transducer structure and in the environment (e.g., air).

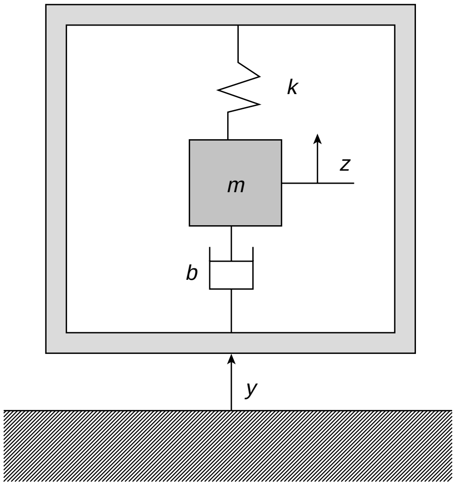


Figure 4. Schematic analysis for mechanical movement of a power generator with external vibration.

Electromagnet Conversion. Motion between the inductor and the permanent magnet induces an electromagnetic current in the inductor coil, as shown in Fig. 5 (25). The induced voltage, V, in the coil is given by equation 4.

$$V = NBl\frac{\omega_n}{2\zeta} \tag{4}$$

where N is the number of turns in the coil, B is the strength of the magnetic field, l is the length of coil, and z is the displacement of the magnet in the coil. This type of generator was fabricated with laser micromachining by Ching et al. (27) in 1 $\mathrm{cm^3}$ volume and generated a 4.4 V peak-to-peak with a maximum rms power of 830 µW. Glynne-Jones et al. (44) at the University of Southampton, derived 157 µW on average new car engine. Perpetuum Ltd., a spin-off company from the University of Southampton, produced an electromechanical microgenerator. That generated up to 4 mW and its operation frequency was 30–350 Hz. The vibration amplitude was 200 µm with 60–110 Hz and demonstrated a wireless temperature sensor transmitter system. This result showed electromagnetic conversion is feasible in low frequency vibration and is promising if it is compatible with silicon micromachining.

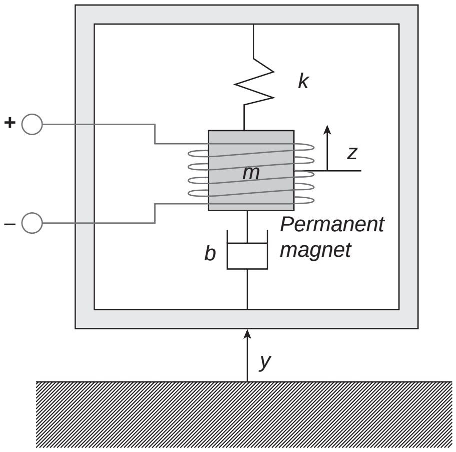


Figure 5. Schematic of an electromagnetic conversion device.

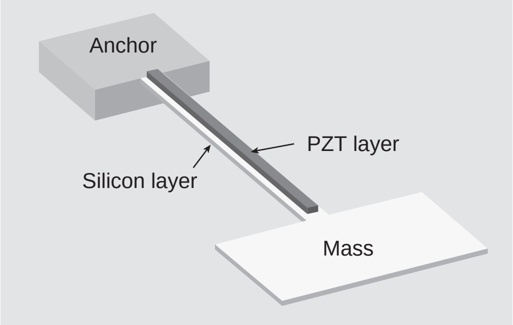


Figure 6. Structure of a piezoelectric cantilever for power generation.

Piezoelectric Conversion. The piezoelectric effect states that the deformation in the material produces an electrical charge due to the separation of charge within crystal structures. The most widely used piezoelectric material is PZT (lead zirconate titanate) in ceramic materials and PVDF [poly (vinylidene fluoride)] in polymers. Several groups are studying the piezoelectric cantilever with seismic mass (Fig. 6). The constitutive equations of piezoelectric materials are expressed in equation 5.

$$\sigma = Y(\delta - d_{31}E) \qquad D = d_{31}\sigma + \varepsilon E \tag{5}$$

where σ is the mechanical stress, δ is the mechanical strain, Y is Young's modulus, d_{31} is the piezoelectric strain coefficient, D is the charge density, E is the electric field, and ε is the dielectric constant of the piezoelectric material. The piezoelectric coefficient links the mechanical stress–strain to the electrical charge equation. If the circuit is open ($D = 0$), the voltage across the piezoelectric layer is described in equation 6.

$$V = \frac{-d_{31}\sigma}{\varepsilon} t_{\text{piezo}} \tag{6}$$

where t_{piezo} is the thickness of the piezoelectric layer. The charge collected on the electrode is integrated on the area of the surface with no load condition as in equation 7

$$Q = \int D\,dA = \int d_{31}\sigma\,dA \tag{7}$$

When the impedance in the load circuit is pure resistance, the time-averaged power can be derived in Ref. 17 with the geometry of a cantilever. White and co-workers (16) presented a thick-film PZT generator, and the maximum power is ~2 μW. According to the analysis of Lu et al. (17), a 5 mm long PZT cantilever can generate >100 μW with the amplitude of >20 μm at ~3 kHz resonance. Roundy (15) demonstrated a piezoelectric converter of $1\,\text{cm}^3$ in volume, that is 1.75 cm in length. It generated 200 μW and is driven with vibrations of $2.25\ \text{m}\cdot\text{s}^{-2}$ at 120 Hz. A microfabricated PZT cantilever generator driven by a bubble was studied by Kang et al. (44), and a few picowatt was generated with one tiny cantilever at 30 Hz and tens of μW is expected on a $1\,\text{cm}^2$ surface (46). If the design, material, and fabrication are optimized, piezoelectric powergeneration will produce hundreds of microwatts with a volume of $1\,\text{cm}^3$.

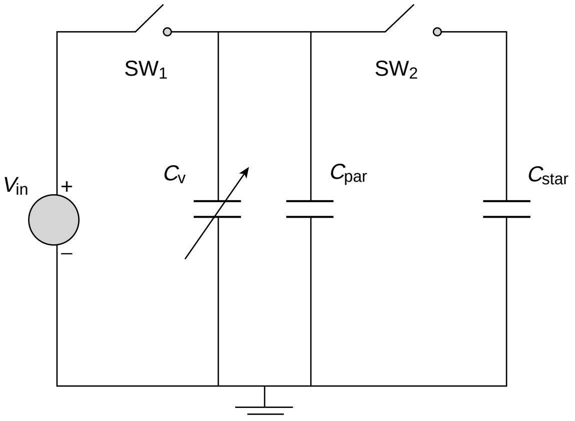


Figure 7. Circuit representation for an electrostatic converter (15).

Electrostatic Conversion. The electrical energy in a capacitor is given in equation 8.

$$E = \frac{1}{2}QV = \frac{1}{2}C_{\text{v}}V^2 = \frac{1}{2}\frac{Q^2}{C_{\text{v}}} \tag{8}$$

When the charge, Q, is constant, if the variable capacitance, C_{v}, is decreased the total energy E in the capacitor will increase. The MEMS structure can change the capacitance C_{v} with an external vibration, and stored energy in the capacitor transfers to energy storage. In the beginning, the external power source V_{in} initiates the charging process as in Fig. 7 (15). When C_{v} is maximum, SW1 is closed and the variable capacitance C_{v} is charged. While vibration changes capacitance C_{v}, all switched are open. When C_{v} reaches a minimum, SW2 is turned on and the energy in C_{v} is transferred to a storage capacitance C_{stor}. The disadvantage of electrostatic conversion is that it needs an external voltage source and switching circuit. The voltage across the storing capacitance is given in equation 9.

$$E_{\text{stor}} = \frac{1}{2}(C_{\max} - C_{\min})V_{\max}V_{\text{in}} \qquad (\text{Ref. } 47) \tag{9}$$

where $V_{\max}$ is the maximum voltage across the capacitor C_{v}. Switching the circuit is realized using a diode and field effect transistor (FET) switch. Meninger et al. (47) made a comb-type variable capacitance with an in-plane overlap type with a 7 μm gap and a 500 μm depth using the 0.6 μm CMOS process. They produced a power of 8 μW with an ultralow power delay locked loop (DLL)-based system. Miao et al. (23) reported an out-of-plane variable capacitor with a gap closing type that varies from 100 pF to 1 pF. A periodic voltage output of 2.3 kV (10 Hz) was generated when the charging voltage was 26 V, which implies that a power of 24 μW ($2.4\ \mu\text{J}\cdot\text{cycle}^{-1}$) can be produced. Mitchenson et al. analyzed architectures for vibration-driven micropower generators (26) and they fabricated a prototype of an electrostatic power generator producing $250\ \text{V}\cdot\text{cycle}^{-1}$ that corresponds to $0.3\ \mu\text{J}\cdot\text{cycle}^{-1}$ (22). Other studies demonstrated polymer capacitor (24) and a liquid rotor power generator with a variable permittivity producing 10 μW (19). Recent developments in electrostatic generators

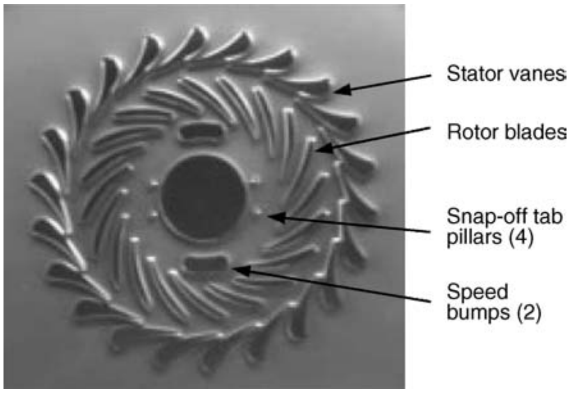


Figure 8. SEM of microturbine of MIT (52).

demonstrated that it is feasible and manufacturable with MEMS and that it has the advantage of a low frequency application like human body movement. However, the generated voltage is very high and should be managed for the implant application. Since it is compatible with the CMOS process and the variable capacitor is a well-established MEMS device, it is very promising as an integrated power generator with a sensor and transmitter.

Microheat Engine

A microheat engine is hard to be implanted in the human body, but it is promising as an external power source for portable medical equipment. A tiny internal combustion engine, made by precise machining, such as electrical discharge machining (EDM) or MEMS, may have a much higher density than a primary battery since the energy density of fossil fuel is $\sim 45\ \mathrm{MJ \cdot kg^{-1}}$, while that of Li ion batteries are at most $0.5\ \mathrm{MJ \cdot kg^{-1}}$ (48). Microturbine by EDM (48), Microrotors by deep reactive ion etching (DRIE) (49), heat engine (42) and reciprocating devices (50) was reported for electric power generation.

Several groups are working on a microheat engine, since fossil fuel offers a much higher energy density. Since the generated power is expected to generate 10–20 W, it is suitable for high power application. The Stirling engine (51), the reciprocal combustion engine (50), the Wankel motors, and gas turbines are reported. One of the first microengine projects was started at MIT (52) and the microfabricated turbine with a 4.2 mm diameter was illustrated in Fig. 8. Massachusetts Institute of Technology has been working on making microheat engines with a turbo charger and Georgia tech is collaborating with MIT on a magnetic generator (53). Allen and co-workers (54) at Georgia Institute of Technology generated a direct current (dc) electric power of 1.1 W with microfabricated windings at 120,000 rpm, although it was not integrated with a heat engine. Peirs et al. (48) made a microturbine by EDM and that tested to speeds of 160,000 rpm and produced a mechanical power of 28 W and an electrical power of 16 W. A miniaturized heat engine is still in the initial stage and no demonstration of power generation using a heat engine was reported. The heat engine would be very useful for high power applications such as portable analytical equipment. Another scheme of a heat engine is thermophotovoltaic power generation (30,31). They converted the heat radiation in a SiC microcombustor to electric energy using photovoltaic cells, which is $<1\ \mathrm{cm^2}$ and produced a power of 1.02 W with 2.28 V.

CONCLUSION

This article reviews micropower devices for medical implantable devices and portable medical device. Since the micropower devices have their own characteristics, there is no winner among them. When a selecting a micropower system for a specific application, one should consider the energy capacity, power, volume, voltage, and compatibility of fabrication with microelectronic device.

Currently, primary or secondary batteries with external power transmission are a main power storage for a low power implanted device. As the application of implant devices is diversified and requires a high power output and longer lifetime, new battery like fuel cell, or biofuel will replace conventional battery system in the future. Furthermore, when power transmission through the skin is impossible due to the attenuation of transmission, integrated power generation device will be an alternative.

Most micropower generators are still in their infancy and they need much more study to be implemented in implant device. Research results on micropower generators showed only the feasibility of concept and their power is much less than the requirement. Hybrid micropower supplies (55) or integrated power system would be strong candidates for long battery life applications. The promising application of active implant device will be glucose sensor and the artificial pancreas for the treatment of diabetes and the ubiquitous bio- or environmental sensor network. Since there is strong demand in the market, it is believed that micropower system will be available in near future.

BIBLIOGRAPHY

1. Soykan O. Power sources for implantable medical devices. Business briefing: Medical device manufacturing and technology. 2002, p 76–79.
2. Mallela VS, Ilankumaran V, Rao NS. Trends in Cardiac Pacemaker Batteries. Indian Pacing Electrophysiol J 2004;4: 201–212.
3. Zhong LC, Shah R, Guo C, Rabaey J. An Ultra-Low Power and Distributed Access Protocol for Broadband Wireless Sensor Networks. Presented at IEEE Broadband Wireless Summit, Las Vegas (NV); 2001.
4. David Culler DE, Srivastava M. Overview of Sensor Networks. IEEE Comput Special Issue Sensor Networks 2004; 41–49.
5. Cardenas-Valencia A, et al. A microfluidic galvanic cell as an on-chip power source. Sensors Actuator B 2003;95:406–413.
6. Sammoura F, Lee Kb, Lin L. Water-activated disposable and long shelf life microbatteries. Sensors Actuators A 2004;111: 79–86.
7. Dyer CK. Fuel cells for portable applications. J Power Sources 2002;106:31–34.
8. Chan SH, Nguyen N-T, Xia Z, Wu Z. Development of a polymeric micro fuel cell containing laser-micromachined flow channels. J Micromech Microeng 2005;15:231–236.

9. Ferrigno R, et al. Membraneless Vanadium Redox Fuel Cell uisng Laminar Flow. J Am Chem Soc 2002;124: 12930–12931.
10. Shah K, Shin WC, Besser RS. Novel microfabrication approaches for directly patterning PEM fuel cell membranes. J Power Sources 2003;123:172–181.
11. Yu J, Cheng P, Maa Z, Yi B. Fabrication of a miniature twin-fuel-cell on silicon wafer. Electrochim Acta 2003;48: 1537–1541.
12. Wainright JS, Saniell RF, Liu CC, Lit M. Microfabricated fuel cells. Electrochem Acta 2003;48:2869–2877.
13. Barton SC, Gallaway J, Atanassov P. Enzymatic Biofuel Cells for Implantable and Microscale Devices. Chem Rev 2004;104: 4867–4886.
14. Heller A. Miniature biofuel cells. Phys Chem Chem Phys 2004;6:209–216.
15. Roundy SJ. Energy Scavenging for Wireless Sensor Nodes with a Focus on Vibration to Electricity Conversion. Mechanical Engineering. Berkeley: The University of California; 2003 p 287.
16. Glynne-Jones P, Beeby SP, White NM. Towards a piezoelectric vibration-powred microgenerator. IRR Proc-Sci Meas Technol 2001;148:68–72.
17. Lu F, Lee HP, Lim SP. Modeling and analysis of micro piezoelectric power generators for micro-electromechanical-systems applications. Smart Mat and Structure 2004;13: 57–63.
18. Shenck N, Paradiso JA. Energy Scavenging with Shoe-mounted piezoelectronics. IEEE Micro 2001;21:30–42.
19. Boland JS, Messenger JDM, Lo HW, Tai YC. Arrayed liquid rotor electret power generation. Presented at IEEE MEMS 2005, Miami(FL) 2005.
20. Roundy S, Steingart D, et al. Power Sources for Wireless Networks. Presented at Proc. 1st European Workshop on Wireless Sensor Networks (EWSN'04), Berlin(Germany). 2004.
21. Roundy S, Wright PK, Rabaey J. A study of low level vibrations as a power source for wireless sensor nodes. Comp Commun 2003;26:1131–1144.
22. Mitcheson PD, et al. MEMS electrostatic micropower generator for low frequency operation. Sensors and Actuator A 2004;115:523–529.
23. Miao P, Holmes AS, Yeatman EM, Green TC. Micro-Machined Variable Capacitors for Power Generaton. Presented at Electrostatics'03, Edinburgh(UK). 2003.
24. Arakawa Y, Suzuki Y, Kasagi N. Micro seismic power generator using electret polymer film. Presented at The Fourth International workshop on Micro and Nanotechnology for power generation and energy conversion applications Power MEMS 2004, Kyoto(Japan). 2004.
25. Li WJ, et al. A micromachined vibration-induced power generator for low power sensors of robotic systems. Presented at World Automation Congress: 8th International Symposium on Robotics with Applications, Hawaii 2000.
26. Mitcheson PD, Green TC, Yeatman EM, Holmes AS. Architectures for Vibration-Driven Micropower Generators. J Microelectomech Systems 2004;13:429–440.
27. Ching NNH, et al. A laser-micromachined multi-modal resonating power transducer for wireless sensing systems. Sensors Actuator A 2002;97:685–690.
28. Qu W, Plotner M, Fischer W-J. Microfabrication of thermoelectric generators on flexible foil substrates as a power source for autonomous microsystems. J Micromechan Microeng 2001; 11:146–152.
29. Wang W, Jia F, Huang Q, Zhang J. A new type of low power thermoelectric micro-generator fabricated by nanowire array thermoelectric material. Proc 22nd Int Conf Thermoelectrics 2003;682–684.
30. Wenming Y, et al. Effect of wall thickness of micro-combustor on the performance of micro-thermophotovoltaic power generators. Sensors Actuator A; in press, 2005.
31. Yang WM, et al. A prototype microthermophotovoltaic power generator. Appl Phys Lett 2004;84:3864–3866.
32. Pescovitz D. The power of small tech. Smalltimes 2002; 2.
33. Lee SJ, et al. Design and fabrication of a micro fuel cell array with flip-flop interconnection. J Power Sources 2002;112: 410–418.
34. Chen T, et al. A Miniature Biofuel Cell. J Am Chem Soc 2001;123:8630–8631.
35. Moore CM, Minteer SD, Martin RS. Microchip-based ethanol/oxygen biofuel cell. Lab Chip 2005;5:218–225.
36. Bates JB, et al. Thin-film lithium and lithium-ion batteries. Solid State Ionics 2000;135:33–45.
37. Pique A, et al. Rapid prototyping of micropower sources by laser direct write. Appl Phys A Mat Sci Proc 2004;79:783–786.
38. Lee KB, Lin L. Electrolyte based on-demand disposable microbattery. Presented at IEEE MEMS 2002, Las Vega. 2002.
39. Applied Digital solutions, www.adsl.com.
40. Biophan technologies. Available at www.biophan.com/biothermal.php.
41. MacDonald SG, Biothermal power source for implantable devices. US Patent 6,640,137, 2003.
42. Whalen S, et al. Design, Fabrication and testing of the P3 micro heat engine. Sensors Actuator A 2003;104:290–298.
43. Williams CB, Yates RB. Analysis of a micro-electric generator for microsystems. Sensors Actuator A 1996;52:8–11.
44. Glynne-Jones P, et al. An electromagnetic, vibration-powered generator for intelligent sensor systems. Sensors Actuators A 2004;110:344–349.
45. Kang J-Y, Kim H-J, Kim J-S, Kim T-S. Optimal design of piezoelectric cantilever for a micro power generator with microbubble. Presented at Microtechnologies in Medicine & Biology 2nd Annual International IEEE-EMB Special Topic Conference, Madison (WI). 2002.
46. Kang JY, Kim JS, Kim HY, Kim TS. Micro Power Generator with Piezoelectric Cantilever Driven By Micro Bubble. Sensors Actuator A submitted, 2005.
47. Meninger S, et al. Vibration-to-Electric Energy Conversion. IEEE Trans VLSI Systems 2001;9:64–76.
48. Peirs J, Reynaerts D, Verplaetsen F. A microturbine for electric power generation. Sensors Actuator A 2004;113: 86–93.
49. Miki N, Teo CJ, Ho LC, Zhang X. Enhancement of rotordynamic performance of high-speed micro-rotors for power MEMS applications by precision deep reactive ion etching. Sensors Actuator A 2003;104:263–267.
50. Lee DH, et al. Fabrication and test of a MEMS combustor and reciprocating device. J Micromech 2002;12:26–34.
51. Backhaus S, Swift GW. A thermoacustic Stirling heat engine. Nature (London) 1999;399:335–338.
52. Frechette LG, et al. Demonstration of a microfabricated high-speed turbine supported on gas bearings. Presented at Solid-state sensors and actuator workshop, Hilton head island, (SC). 2000.
53. Jacobson SA, et al. Progress toward a icrofabricated gas turbine generator for soldier portable power applications. Presented at 24th Army Science Conference, Orlando (FL). 2004.
54. Das S, et al. Multi-Watt electric power from a microfabricated permanent magnet generator. Presented at IEEE MEMS 2005, Miami(FL). 2005.
55. Harb J, LaFollete R, Selfridge R, Howell L. Microbatteries for self-sustained hybrid micropower supplies. J Power Sources 2002;104:46–51.

See also BIOTELEMETRY; COMMUNICATION DEVICES; MICROFLUIDICS.

MICROSCOPY AND SPECTROSCOPY, NEAR-FIELD

Miodrag Micic
MP Biomedicals LLC
Irvine, California

INTRODUCTION

The explosion of knowledge in life sciences is enabled by the ability to visualize beyond the capability of the human eye and by the capability to identify chemical compositions and the structure of matter. This was enabled by Levenhook's discovery of the new device for looking at the world of the small: the microscope. He found that by combining two lenses, it was possible to see much smaller objects than by the naked eye alone. This led to his subsequent discovery of the cell, which has spurred an explosion of knowledge in life science and medicine that continues at a dramatic rate of growth even today. Even 400 years after the Levenhook discovery, one of the first tools of choice for the visualization of small objects is the optical microscopy. What is known to a lesser extent is that the microscopic histochemical studies (i.e., staining of the tissues) with tissues and organelle-specific dyes in the late 1800s, initiated the modern pharmaceutical industry, when chemists and histologists alike envisioned an opportunity to selectively inhibit or kill pathogens by organic molecules in the same way organic dyes selectively label certain types of tissues, cells, and organelles. While the optical microscopy methods were able to uncover morphology and the structure and nature of the cells, they were faced with the ultimate physical limit of magnification, which is dictated by the spatial resolution limited by the diffraction limit. This limit was approximately the size of half of the wavelength of light used to perform the imaging.

The breakthrough in imaging small structures and further understanding the machinery of life and cell biology comes with the application of the deBroglie's postulate of particle-wave equality in order to use the electron beam with a shorter associated wavelength as an investigative imaging probe. Ruska's development of the first transmission electron microscope in late 1930s and the development of scanning electronic microscopies in the 1950s, opened the door to detailed investigations of the organization of subcellular assemblies, viruses, and even imaging of the individual biomolecules, at a resolution far beyond the diffraction limit of visible light. The rapid advances in the tools and techniques of ultramicroscopy, especially of scanning probe microscopies, which for the first time enabled routine molecular imaging, greatly contribute to enabling completely new multidisciplines, like nanoscience and nanotechnology, as well as an opening a door for an entirely new way of looking into the machinery of life.

While scanning probe microscopy in the 1990s allowed imaging at the unprecedented resolution of the unaltered samples; in general, it lacked the ability to uniquely identify the chemical composition of samples or unveil their physicochemical properties. For the investigation of chemical structures and fingerprinting the material composition, the tool of choice is optical spectroscopy. However, the problem with classical optical spectroscopic techniques is that it provides average, bulk results with no specific information linking certain morphological features with spectra. This can ultimately identify chemical composition and/or physicochemical properties. The ability of doing molecular fingerprinting and, in a raster pattern, subsequent molecular specific imaging at the nanoscale with the spatial resolution of modern ultramicroscopy techniques is the holy grail for many aspects of today's life sciences disciplines.

This goal is partially fulfilled with electron microscopy combined with energy-dispersive X-ray analysis (EDX), wherein semiquantitatively, it is possible to associate topographical structures with elemental composition. However, for most of the problems in life and materials sciences, simple knowledge of elemental composition is not sufficient, as it is necessary to identify molecular structure. Plus, the electron microscopy is, in most cases, a destructive method of analysis, since the sample needs to be prepared to be vacuum compatible, and be either electrically conductive, in the case of scanning electron microscopy, or have contrasts with heavier metals, in the case of transmission or scanning transmission electron microscopy. Furthermore, the physics of generating characteristic X rays (i.e., the minimum size of the excitation volume from which the signal is emerging) is in the tens of micrometers, thereby limiting elemental compositional analysis with spatial resolution only for the large structure in the tens-of-microns-sized range. The ideal technique will be one that will allow imaging of the unaltered sample, in a way similar to the way atomic force microscopy (AFM) allows, while at the same time providing a way for spectroscopic identification of the chemical structure.

There are several techniques currently in their infancy that promise spectroscopic probing with electromagnetic spectroscopic information carrier signals imposed over topography. They are near-field scanning optical microscopy, microthermal analysis, scanning nuclear magnetic resonance (NMR) microscopy, and scanning electron paramagnetic resonance (EPR) microscopy.

However, for solving any of the practical problems, it will be of great benefit that the ultrastructure's information probes are photons of visible, and near-infrared (IR) and ultraviolet (UV) light, as a great deal of both morphological (based on the photon's position and intensity/count) and compositional (based on adsorption, fluorescence, Raman shift, etc.) information can be simultaneously obtained as optical microscopy relies on light as an information carrier. This is due to the fact that the same photons, which are in standard imaging configurations used to generate images, carry much more information on composition and the various physical and chemical properties of the observed spot on the sample that can be extracted through different spectroscopic methods.

The technique that has evolved over the last decade and promises to fulfill the above-goals at the nanoscale level, is near-field scanning optical microscopy (NSOM or SNOM), which effectively breaks the physical limits imposed by the optical-diffraction-limited resolution by using the near-field evanescent waves and scanning mechanisms similar to those in the scanning probe microscopies.

THEORETICAL PRINCIPLES OF NSOM MICROSCOPY

Abbe's equation (Eq. 1) describes the resolution of the classical, far-field optics, (i.e., the minimum separations between two points that can be distinguished) (1). From this equation it is easy to conclude that the maximum attainable resolution in the far field is $\sim \frac{1}{2}$ of the applied wavelength, which means that the best optical microscope cannot be used to visualize details smaller than 200–400 nm.

$$p_{\min} = \frac{\lambda}{2n \sin \alpha'} \quad (1)$$

In Abbe's equation, n is the diffraction of the imaging index and α' is the aperture angle in the medium. While Abbe's equation describes the limiting resolution in the world of conventional optics, the Fourier optics approach can provide us with the same conclusion. Following Abbe's principle, Rayleigh (2) derived that the objects in a lens system in the far optical field are resolved only when the maximum of one pattern coincides with the first minimum of the neighboring features. What resulted was the discovery that the resolution criteria that describe the maximum resolution of the optical system based on the size of the numeric aperture was

$$d = 0.61\,\lambda/\mathrm{NA} \quad (2)$$

wherein λ is the applied wavelength and NA is the numerical aperture of the lens, again bringing the maximum theoretical resolution to $\sim$ 200 nm.

A similar observation can be derived from the Fourier formalism in optics. In Fourier optics, the information content embedded in the spatial frequency f, in the case when f is higher then $1/\lambda$, decays rapidly toward zero from the object and thereby, no data on the subwavelength features can be efficiently collected with standard far-field optics. However, it is well known that it is possible to receive an electromagnetic signal with antenna that is smaller in size than the wavelength. The solution of the Maxwellian equations that govern the behavior of electromagnetic radiation differs in the distance smaller than the wavelength than in the distance, much larger than the considered wavelengths. When the waves propagate within the distance much smaller than its wavelength, such situation is called the near field. The pragmatic definition of near-field optics will be a division of optics that deal with the elements of the subwavelength features scales, which are intended for passing the light through, from or near, to another element with subwavelength features positioned within the subwavelength distances. The spatial resolution in near-field optics depends on the feature's size and is limited to about one-half of the aperture size. Furthermore, the near-field system must be considered as a complete system consisting of two features (probe and sample in the case of microscopy) and the resolution of the system will be dependent on both sample and probe. Thus, it is impossible to speak of the unique or standard near-field resolution, as is done with a far-field instrument.

NEAR-FIELD IMAGING EQUIPMENT

The near-field scanning optical microscopy or scanning near-field optical microscopy (NSOM or SNOM) is a technique that enables users to work with standard optical tools that are integrated with scanning probe microscope (SPM) technology to obtain the optical image at a resolution in the range of tens of nanometers. This is quite comparable with the resolution of scanning electron or SPM. The integration of scanning probe and optical methods allows for the collection of optical information at resolutions well below the optical diffraction limit, which overlap real topography information obtained by scanning probe feedback. For spectroscopy applications, NSOM offers the potential for characterizing the spectroscopic signature of material on a submicron-to-nanometer scale, thereby affording new insights into nanoscopic structure and composition.

The principles of NSOM microscopy were theoretically founded by Synge in 1929 (2), and in his subsequent paper he described an imaginary device that closely resembles today's NSOM setup (3), including the use of piezoactuators. Due to technical difficulties at the time to implement such a device, the idea was forgotten until theoretician O'Keefe rediscovered the idea in 1956 (4). The first practical demonstration of the imaging of a structure smaller than the one-sixtieth of the applied wavelength was done in 1971 using the microwave in near-field scanning over the grid (5). The basic idea behind this method was to create evanescence, a standing wave, by light diffraction through an aperture that was much smaller than the wavelength, and then to use this evanescent light source to scan a sample in a raster-scan pattern in close proximity, and collect transmitted or reflected signal in the far field. The first demonstration of near-field optical imaging was implemented independently by Pohl (6) in 1982 and Lewis groups (7) in 1983, and described as an optical stethoscopy, in what is now considered the beginning of NSOM microscopy. The method grew rapidly during the 1990s and the trend is continuing to this day, as described in recent reviews (8–12). Furthermore, several companies are offering commercial instruments (13–16) that enable ordinary users, who are not inclined toward the instrumentation development, to apply NSOM in solving their research problems.

There are two fundamentally different ways to achieve near-field optical imaging. They are apertured-base and apertureless NSOMs with their principles of operation depicted in Fig. 1a and b (17). In the case of the apertured NSOM (Fig. 1a), the light passes through an aperture that is much smaller than the wavelength of applied light, and ultimately the resolution is defined by the size of the aperture. In the case of the apertureless NSOM (Fig. 1b), a sharp metallic tip is irradiated by a laser perpendicular (or as close as possible) to the tip along the axis. The irradiation excites the plasmons on the metallic surface of the tip and the field is concentrated by combining antenna and plasmonic effects at the top of the tip. The resolution of apertureless NSOM is thus defined by the size of the near-field excitation formed at the apex of the metallic tip, and is determined by tip sharpness, tip materials, and real and imaginary parts of the refraction index of the used metal.

The practical advantage of apertured NSOM is in its easy implementation. While the advantage of the apertureless

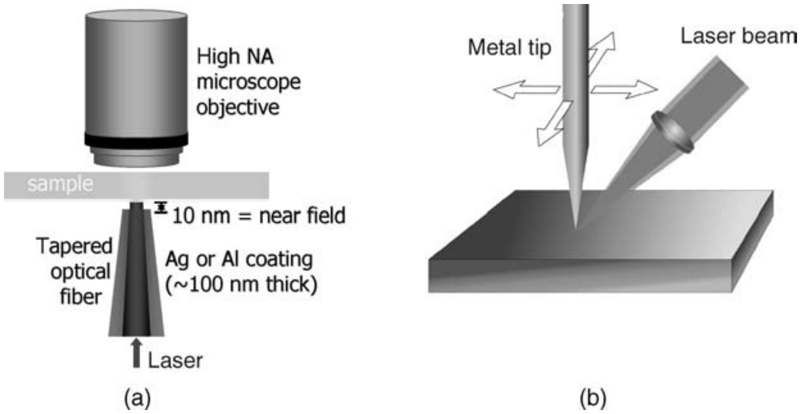


Figure 1. Two major principles of achieving the near-field scanning optical microscopy: (a) the aperture-based NSOM; (b) scatter-field, apertureless NSOM.

NSOM is theoretically a higher achievable resolution and possibly higher field strength, the technical difficulties in implementing the apertureless setup have not permitted those advantages to be realized. Thus, as of this writing, all of the existing commercial instruments are based on the apertured NSOM approach.

Regardless of the type of NSOM, the experimental setup always consists of a piezo X–Y scanner, whose role is to execute the raster-scan pattern scanning of the sample by the near-field probe; a Z piezo, whose role is to modulate the image by keeping the sample-tip distance; an optional, but for most cases necessary, noncontact modulation element, such as the tuning-fork for shear force feedback, or a piezo- or electromagnetic oscillator for AFM-like noncontact feedback; a near-field probe, which can be aperture or sharp tip; a laser light source; a far-field optical signal collection system, sample and sample holder; a system for coarse probe approach and sample–probe alignment; a scanner controller and computer for the image acquisition and reconstruction; and a vibration isolation system. In this manner, the NSOM most closely resembles the mechanism of the AFM, and almost all of the existing commercial setups, as well as many of the in-house, lab-made NSOMs, share these common components with the AFM and more general SPM platforms.

Piezo Scanner

The piezo scanner principle of work is based on the piezo effect, which is reversible internal stress induction within the crystal when exposed to the electric field (18). This stress induces crystal expansion. The piezo scanner in the NSOM is a more critical part than in the standard AFM. Ideally, it should be the perfect closed-loop scanner, due to stringent requirements of keeping the probe in a particular place during the raster scan, in order to achieve sufficient optical signal/noise ratio. Figure 2 depicts typical scanner configurations. The scanners are usually implemented in the form of the stacked piezo crystals (Fig. 2a), tube scanners (Fig. 2b), and bimorph (Fig. 2c) (19). Furthermore, scanners are often grouped into the so-called tripod configuration. The same material used for the SPM scanner, lead-zirconate-titanate ceramic (commonly referred as a PZT ceramic), is commonly used for the NSOM piezo scanner. The typical piezo-electric constant for PZT materials is about $-1.7\,\mathrm{V\cdot nm^{-1}}$. However, in order to practically achieve the linearity over the whole range of scan, it is necessary to calibrate each individual scanner periodically to compensate for crystal nonlinearity, creeping, and drift. Those effects are further minimized by using active, real-time feedback, which can be implemented either through some form of the strain gauge, capacitance, or by optical means. The active feedback adjusts the voltage applied to the scanner to keep linearity and to secure the probe above the scanning position within the raster scan.

Optical Signal Acquisition System

The optical signal collection system is made up of optical and optoelectronic parts. The optical portion usually consists of the far-field microscope objective with a high NA lens. The tip-sample working distance, as well as the sample thickness and sample holder accessibility, define the maximum NA of the objective that can be used. Oil immersion objectives are used to enhance the NA by many times. Besides the objective, the collection system may contain filters, notch-filters polarizers, and beam splitters, depending on the particular configuration and imaging mode. The optoelectronic part of the collection system converts optical information to an electrical signal for further processing. It is usually either a highly sensitive photomultiplier tube (PMT), or, for ultimate sensitivity and single-photon counting, an avalanche photodiode detector (APD) array. For the PMT tube, the output signal can be either voltage or counts, and for the APD, it is only TTL (transistor–transistor logic) counts. The high sensitivity PMT tubes can satisfy most of the imaging requirements, however, for extremely weak signals, such as in single-molecular imaging or single-molecular spectroscopy, an APD detector is more desirable. Due care does need to be paid when using the APD detector, as overexposing the detector can damage it in an extremely short period of time. For the purpose of correlated experiments, many detectors are attached to the system. In the case of spectroscopy applications, the most commonly used dispersive detector is a highly sensitive CCD imaging camera, either solid-state or liquid-gas cooled. However, many setups use the other, more economical, wavelength or energy-dispersive detection systems, which are based on filters, prisms, or gratings in conjunction with either a PMT or APD detector.

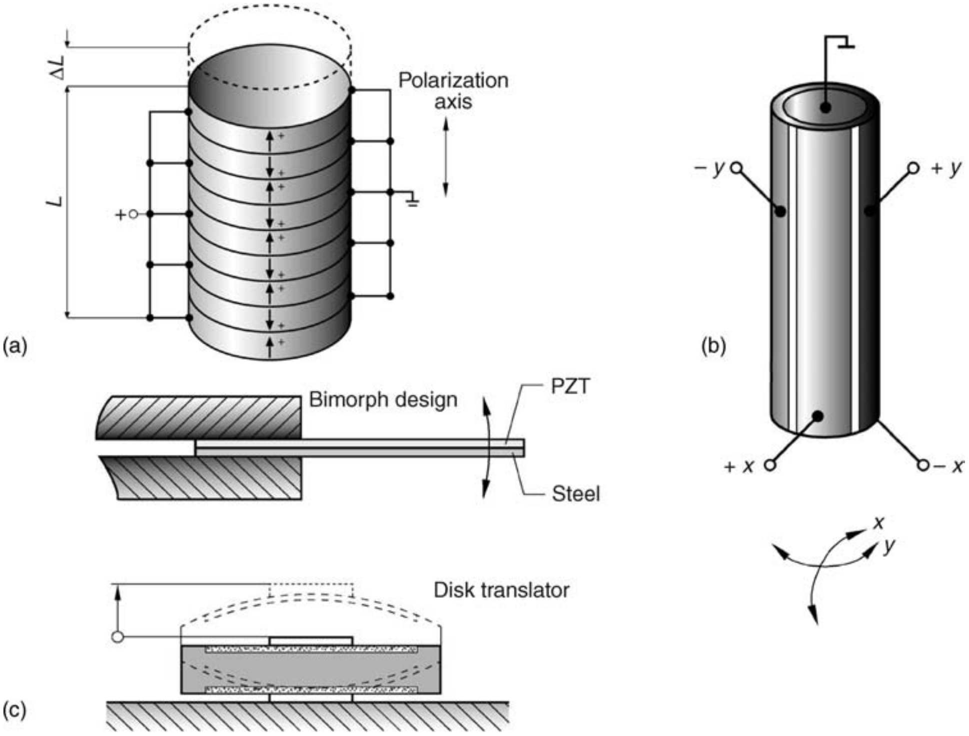


Figure 2. Schematics representation of the piezo actuators: (a) stacked piezo; (b) tube scanner; (c) bimorph scanner. [Courtesy PI (Physik Instru-mente) LP, www.pi.ws.

Light Sources for Near-Field Imaging and Spectroscopy

Illumination sources for near-field microscopy are always lasers. Selection criteria for the laser depends on the desired wavelength(s). The most economical are the solid-state lasers, followed by ion lasers, such as the Ar laser, which can selectively produce multiple wavelengths of light from 450 to 514 nm (17). Besides the two most common types, many setups use liquid lasers as well as optical parametric oscillators to produce specific wavelengths that are not available with a standard laser. The transduction path can consist of mirrors or single-mode optical fibers. The mirror-based path has better throughput, however, it is more complex to adjust and requires periodic readjustment. The optical-fiber-based transduction path has some higher attenuation then the mirror-based path, but it is very convenient for use, especially if it is implemented with the standard FC or similar connectors.

Microscope Head

The NSOM head usually consists of a probe holder, one or more piezo scanners, a system for coarse probe approach, and a case. The head is positioned at the top of the sample holder. The probe holder is directly attached to the *Z* direction piezo. A video system, which shows the image of the approaching tip and substrate, and a software- controlled stepper motor with micrometric screws provide coarse approach of the probe to the sample surface. When the probe tip is brought to close proximity to the sample, the fine approach mechanism is engaged. The fine approach mechanism is essentially a stepped mechanism wherein the probe is brought in a small piezo steps in the vertical direction to close the gap between the tip of the probe and sample substrate. The contact is achieved when the signal indicates the deflection of the probe in AFM-like setups, or, in shear-force mode, when its interaction with the sample reaches the user-prescribed "set point" voltage or current level. The determination of the appropriate set point varies for different samples and systems, and is more a result of art or tacit knowledge than an exact science. If the *Z* piezo is completely extended and contact has not been achieved, the piezo constricts to its neutral position and the stepper motor is activated to bring the probe to the approximate max extension distance of the piezo, and the process is repeated. Besides the *Z* piezo, sometimes the *X*–*Y* scanners can be positioned in the head.

Sample Holder/Stage

The sample holding stage can contain the *X*–*Y* piezo scanner, if it is not in the head. Its moving frame consists of micrometric screw positioners that push the sample holder in the *X*–*Y* direction under the probe, thus allowing sample "pan" operation. These screws can be manually or stepper-motor operated. The sample stage can be stand-alone, or it can be positioned at the top of an inverted, epi-fluorescence microscope. There are many advantages to having the NSOM sitting at the top of the standard inverted microscope. This configuration is able to combine far- and near-field microscopy, exploit the operation familiarity of the inverted fluorescence microscope, and deliver superior images to those acquired via a dedicated, stand-alone NSOM stage. However, a drawback of such a configuration is a larger mechanical circuit with a higher level of vibrational noise than in a dedicated system.

Controller

Controllers for NSOM are usually derived from the AFM/SPM controllers. In all of today's setups, they are digital.

The controller's role is to generate the high voltage signals necessary to feed the piezo scanner and move the probe in the raster scan pattern. It also controls the vertical, Z position of the probe via the PID control-loop model mechanism (proportional-integral-differential), and thus topographically modulates the signal; it maintains the non–contact feedback; it controls the coarse probe approach, and in some instances coarse sample positioning; and it acquires the signals coming from the probe (both optical and topographical) and forwards them to the computer for further processing. The controller usually consists of a series of analog-to-digital and digital-to-analog converters, precision operational amplifiers, and high voltage amplifiers. Due to the complexity of the tasks, often the controller is designed using high end digital signal processors and other high end embedded systems.

The role of the control software is to control the controller, acquire the image, and store it in some of editable and exportable format. Furthermore, the control software almost always possesses image processing capabilities, such as different image filters, Fourier transform, 3D representations and rendering, and so on. All of the commercially available NSOMs share the same software with their AFM/SPM "cousins". Many of the homemade systems, on the other hand, have software modified from existing commercial SPM/AFM controller software, or software that is independently written, as in cases where the users have designed the whole control electronics by themselves. Many times, the results are processed in third party software. For example, for image processing a very popular solution is to use shareware NIH Image software or its PC cousin, Scion Image, and for advanced applications, to use scripts written for IgroPro, LabView, MathLab, or for spectroscopy experiments, WinSpec. Use of higher level software like Igor Pro dramatically reduces development time of applications, as compared to the time required to write the script in C or C++ code.

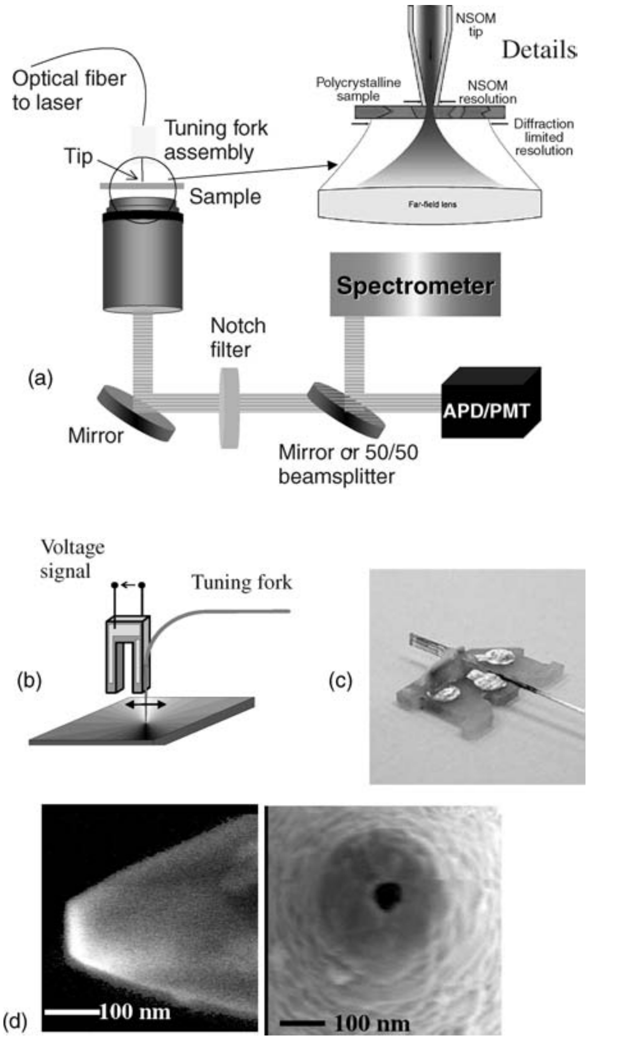


Figure 3. Example of the aperture-based, straight-fiber NSOM: (a) schematics representation of the scanning probe and collection; (b) schematics representation of the tapered fiber NSOM probe, attached to the tuning fork; (c) photography of the probe glued to the tuning fork; (d) SEM image of the end of the tapered, aluminum-coated fiber optic based NSOM probe. (Courtesy of Veeco Instruments Inc., Santa Barbara, CA.)

Apertured NSOM

The principle of the apertured NSOM is to use the aperture as a scanning probe. This is the first (5,6) and up-to-today most commonly implemented NSOM setup. In basic principle, the instrument consists of the XYZ piezo scanner(s) that moves the apertured probe over the raster scan pattern at the controlled aperture-sample height, and a noncontact feedback mechanism, the best-suited being the shear-force based one (20). An aperture in the tens of nanometer size can be formed by the tapering, heating, and pooling process borrowed from biophysics labs, where it is utilized for creating micropipettes (21) or for etching optical fibers (22). Additionally, as an aperture, it is possible to use the hollow cantilever (23).

Schematic representation of the instrument implementation, for both imaging and spectroscopy–hyperspectral imaging, is presented in Fig. 3a, configured for the most common, transmission mode (looking through the sample) operation. The laser light is coming from the optical fiber. Optical fiber is mounted on the tuning fork assembly, which is held onto the Z-piezo scanner and is constricted at the end to a tens of nanometer size range (see insert) and the evanescent field is formed at its aperture, as represented in the figure insert. The light is passing through the sample, interacting with sample matter, and is collected under the sample in the far-field with the high numerical aperture microscopy objective. Such a signal is further subjected to collection and processing. For the hyperspectral imaging or for the spectroscopy or spectral imaging purposes, the signal is first passed through the holographic notch filter, which eliminates excitation light and through the beam splitter is directed to the wavelength-dispersive detector, such as a CCD spectrometer, and to the summary, imaging detector, such as a PMT, or avalanche photodiode detector (APD). In the simplified setup, if the apparatus is used just for the optical imaging, the light is passed directly from the objective into the imaging detector, (i.e. APD or PMT).

Figure 3b schematically represents a typical apertured probe, mounted on the tuning fork. Micrographies at

Fig. 3c and d represent the frontal and lateral view of the metallic-coated, laser-pulled, tapered, fiber-based tip. An evanescent, standing wave is formed at the end of the aperture, and the size of the aperture approximately defines the optical resolution. Light is either brought through the aperture, as in transmission and reflection imaging mode, or collected through the aperture, as in the collection mode. The tip is scanned across the sample in a raster-scan pattern, and for imaging purposes; the signal is collected in the far field, either by sensitive photomultiplier tube, or by sensitive avalanche photo-diode counter.

The three distinctive different modes of operations of the aperture-based NSOM are illustrated in Fig. 4, which also graphically depicts the different kinds of information that can be extracted from the optical signal emanating from the sample. The origin of the optical contrast, as depicted in Fig. 4, can be due to topographic differences (different path length change the adsorption), material birefringence, reflectivity, sample extinction coefficients for the particular excitation wavelength, index of refraction, fluorescence emission properties, nonlinear spectroscopical properties of materials, and mechanical and magnetic stress in the sample. However, at the moment of this writing, for life sciences and biomedical applications, only transmitivitty, reflectivity and fluorescence properties are of significance. The mode that is the most useful for biological applications is the transmission mode or the "looking through" mode, and is most similar to classical biological microscopy In this case, as described above, light is brought through the fiber-based tip, the near field interacts with the sample, and the signal is collected in the far field as it passes through the sample. In this mode, it is possible to do transmission imaging, as well as fluorescence or other wavelength-resolved imaging, by application of adequate filters or wavelength-selective elements. In a reflection mode, which can be described as "looking on the surface mode", the near field interrogates the surface of the sample, and the scattered signal is collected in the far field. This mode allows imaging and spectroscopy of the nontransparent samples; however, the imaging efficiency is much lower than with transmission mode. In a collection mode, the light is passed either through the sample, or illuminated on the sample, and the signal is collected through the fiber in the near field. This mode is very burdensome to use, has low signal collection efficiency, and is used mainly in photonics research.

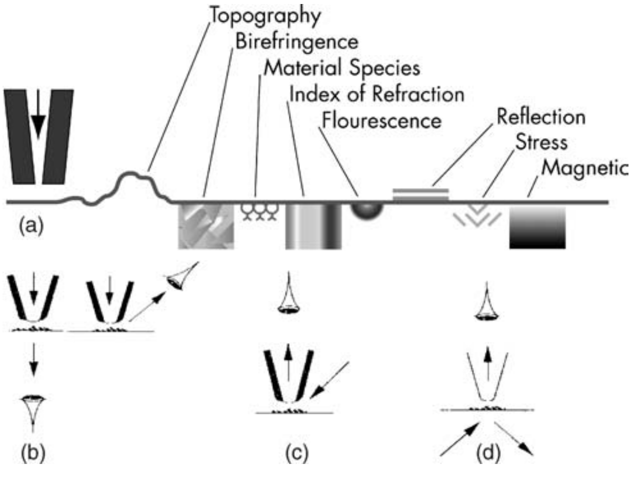


Figure 4. Typical apertured NSOM configuration: (a) illustration of information that is carried and can be extracted from the optical signal; (b) transmission and reflection mode NSOM; (c) collection mode NSOM; (d) total internal reflection or dark field mode. (Figure 3b–d adapted from Ref. 17.)

Besides these three modes, there are more exotic modes of operation, such as combined collection and illumination, where both sample illumination and resulting signal are passed and collected through the same probe; dark field imaging, where the probe tip is in close proximity to a sample that is illuminated from underneath with total internal reflection from the substrate, and wherein the probe acts as a second, tunneling prism. Besides pure optical operation modes, there are also a combined opto-magnetic NSOM, which explores Kerr's effect (24); nano-mass spectroscopy (25), where near field is used for ablation; and optoelectrochemical NSOM (26). The later two modes have a lot of potential applications in physiology with their ability to simultaneously image and record potential at subwavelength resolution.

To secure the probe in a near field, the aperture must be kept in close proximity to the sample with a distance much smaller than the applied wavelength. The fiber is kept at the nanometer-range distance from the sample by means of noncontact feedback. There are several ways of achieving feedback, mainly shear force, AFM-like normal force contact, and noncontact force feedback. The shear-force feedback provides gentler, lateral touching of the sample, thereby reducing the possibility of aperture contamination or tip–aperture mechanical failure.

In shear-force feedback (20), the fiber tip is oscillated laterally to the sample surface. The NSOM tip is rigidly premounted on a quartz tuning fork (Fig. 3b and c), which is a few millimeters in size. The tuning fork is mechanically vibrated at resonance frequency, usually in tens to hundreds of kilohertz, resulting in a few nanometers of lateral motion at the distal end of the NSOM tip. When the tip is in a close lateral proximity to the sample, the resonance frequency of the tip-tuning fork system is disturbed due to electrostatic, van der Waals, hydrogen bonding, and other kinds of attractive and/or repulsive interactions between the tip and the sample. This disturbance is read as an electrical signal that is processed, and the tip is moved accordingly in the vertical direction to achieve its preset resonance frequency, thus keeping the same distance from the sample.

Optical resolution, which is typically achieved by fiber-based apertured NSOM, is in the range of 50 nm, with maximum resolution being in the range of 20 nm. The improvement in tip fabrication procedures and in the control of the tip-sample separation distance will ultimately lead to better resolution. For apertured NSOM, fiberoptic or pipette-based tips are fabricated by constricting the core of the optical fiber to a 50–20 nm diameter. This is achieved by a heating–pulling method (21), wherein the fiber is transversally irradiated by CO_2 laser and simultaneously stretched on the pipette puller until the fiber is broken, or by chemical etching (22) at the phase boundaries using the HF solution with oil on the top. To enhance the efficiency of the light transmission and to avoid the light leakage through the fiber shell, the probes are usually coated with either aluminum or silver. The coating is done by vacuum evaporation, and its role is to prevent light leaking out of the probe. The metallic coating is especially beneficial in the near-field surface-enhanced Raman spectroscopy.

Another way of performing apertured NSOM is by bending the fiber or pipette. In this case, the force feedback can be achieved either by shear force, or by AFM-like normal force in both contact or noncontact mode. The disadvantage of this approach is that such bended fibers are more vulnerable to mechanical failure, and if used for spectroscopy purposes, there may be problems with Raman scattering lines coming from the fiber shell materials.

The other way of achieving apertured NSOM imaging is by using the hollow AFM cantilevers (23). In this kind of setup, the light from the excitation laser is focused on the top aperture on the center of the hollow AFM tip, and the near field is formed at its bottom. The feedback mechanism used therein is the same as in noncontact AFM. Presently, the resolution (in the range of 100 nm) of such hollow-cantilevered-based apertured NSOMs is inferior to that of pulled-fiber-based NSOMs. Another disadvantage of the AFM-like force-feedback setup in NSOM applications is in that the AFM uses a laser beam to follow the bending of the cantilever. In NSOM, when many applications are counting the individual photons, the optical noise introduced by the AFM-like laser-based feedback may be several folds stronger than the signal. Considerable improvement is to be expected with piezo-actuated hollow cantilevers, which will avoid using laser feedback.

Apertureless NSOM

In the apertureless NSOM (Figs. 1b and 5), a sharp metallic tip is irradiated by the laser light orthogonally to the long tip axis, and the near-field excitation is scattered from the tip (27,28). The light scattering from the feature is much smaller than the applied wavelength, which also generates the strong evanescent field. The best strength of the scattering field is achieved if the excitation laser frequency corresponds to the surface-plasmon resonance of the metal from which the tip is made. Incoming beam scattering produces the evanescent field at the tip; however, the physics of the process is a combination of the near-field antenna effects and surface plasmon resonance. The laser induces the plasmons in the tip, which oscillates in parallel to the tip axis and amplifies the evanescent standing wave at the tip apex. The standing wave interacts with the sample and the signal is collected either in transmission or reflection mode in the far field. The tip is scanned across the sample in the same raster-pattern manner as with apertured NSOM. Feedback is provided in either the noncontact AFM manner, preferably with a tuning fork or some other nonlaser based Z-deflection feedback, and the signal collection is modulated by oscillating the probe in a vertical direction in order to avoid static and scattering artifacts. Furthermore, in modulated apertureless NSOM, the signal from the photodetector is also modulated with the same modulation signal source as a tip, in exactly the same frequency and phase (with possibilities of higher harmonics modulation. This is done in order to avoid inbound laser light nonnear-field induced scattering; static-scattering artifacts from sample features and to achieve optical signal acquisition always in a same sample-tip separation distance position.

In order to distinguish between the near-field scattering and inbound laser light, most of the apertureless NSOMs are used mainly for fluorescence, Raman, or for different nonlinear optical phenomena applications. Furthermore, because of the rapid decoy of the scattered field, modulation of the scanning probe, and control of the sample-tip separation distance in the apertureless configuration is much more critical than in the aperture-based NSOM.

Figure 5 is a schematic representation of a typical, homemade, apertureless NSOM setup; in this particular case used for fluorescence and fluorescence-lifetime (FLIM)

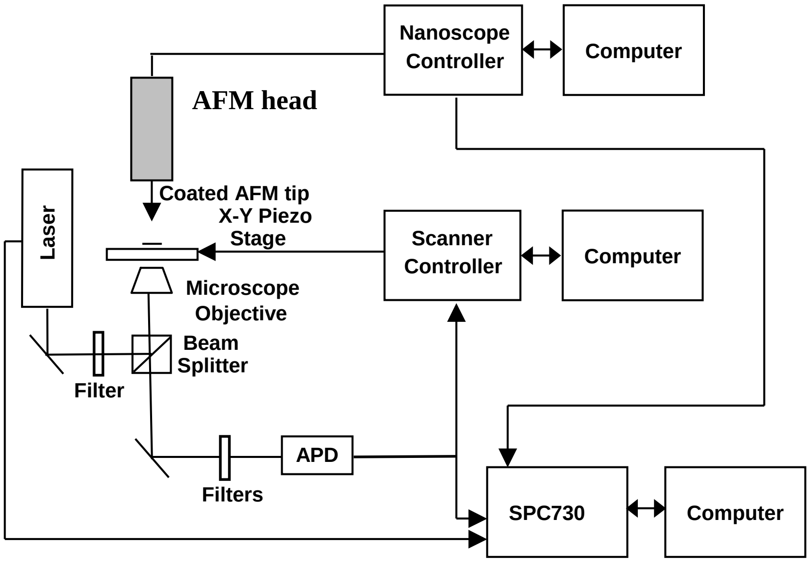


Figure 5. Schematic representation of the example of the apertureless NSOM. (Adapted from Ref. 29.)

near field imaging. It consists of a commercially available AFM head, mounted on the top of the inverted epi-fluorescence microscope, positioned at the optical table. The lateral irradiation of the tip, which is necessary in order to achieve the high intensity evanescent field generation, is produced by offsetting the position of the AFM tip in relationship to the high numerical apex microscope objective. In this particular case, the fluorescence imaging measurements were conducted in an inverted fluorescence-imaging microscope (Nikon Diaphot 300), the excitation light from a mode-locked YAG laser (Coherent Antares) at 532 nm wavelength, 10 ps pulse width, and 76 MHz repetition rate was focused on a diffraction-limited spot through an objective (Nikon 60X NA 1.4), and the emission from the sample was collected by the same objective, in the epi-fluorescence manner. The emission band-pass filters were HQ565/25 and D570/20 (Chroma Technology) to ensure that the excitation light and the feedback laser of AFM (650 nm) were both blocked. The emission was detected by an avalanche photodiode (APD) (Perkin Elmer, SPCM-AQR-15). The background photon counts with AFM feedback on were ~150 Hz. The sample cover slip was mounted on a closed-loop two-dimensional (2D) piezoelectric scanner (Polytec PI, P-731). The AFM (Veeco Instruments Inc, D3100) head and inverted microscope were coupled at an over–under position. The AFM tapping-mode tips used in this work are commercially available Si tips (Digital Instrument, OTESP7) coated with Au and Ag, by sputter coating. Image density of 128 × 128 pixels and scan rate of 1 Hz. As the quenching effect is highly distance dependent, the tip oscillation amplitude was reduced by reducing the driving voltage to the tip as much as possible without sacrificing image quality. Based on the force calibration curve, the tip oscillation amplitude during the imaging was estimated at ~ 30 nm.

The sample-scanning confocal fluorescence image was recorded by a home-built computer control interface that counted the APD signal and raster-scanned the piezoelectric scanner. The fluorescence decay traces were recorded by a time-correlated single photon counting (TCSPC) module (Becker & Hickl SPC730, Germany). The start signal was from the APD and the stop signal was from synchronization of the YAG laser at one-half of the laser repetition rate. For the lifetime imaging mode, the TCSPC module reads the line-synchronization signal of the Digital Instrument Nanoscope IIIa controller to achieve a synchronized recording of the AFM signals and fluorescence signals.

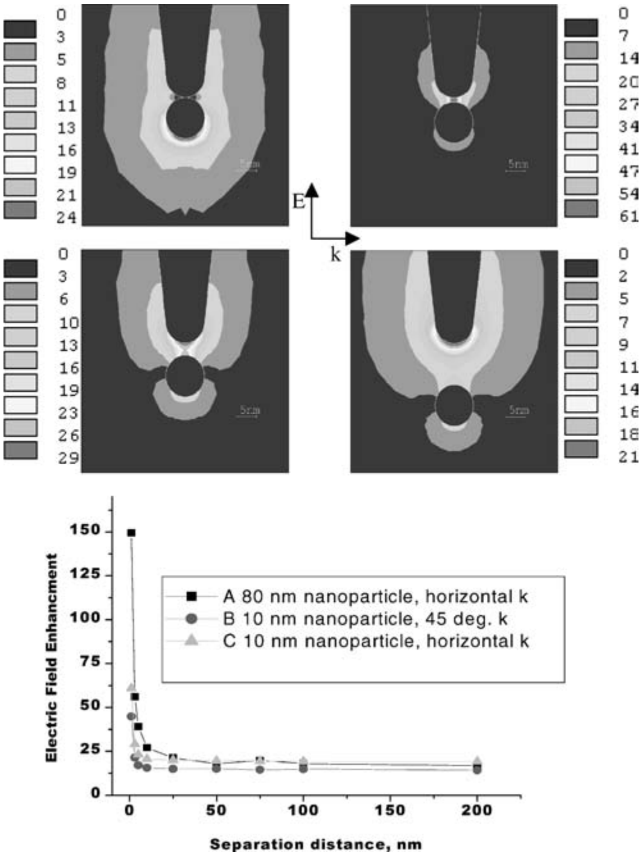


Figure 6. Finite-element methods (FEM) simulation of the electromagnetic field scattering and near field enhancement at the apex of the tip of the apertureless NSOM, and its behavior with change in tip-sample separation distance. (Adapted from Ref. 30.)

Figure 6 depicts the FEM simulation (30) of near-field enhancement around a metallic tip positioned in close proximity to the sample and irradiated with a laser beam. Figure 6 shows the rapid dependence of the near-field excitation on the sample-tip separation, and emphasizes necessity of accurate, sub nanometer sample-tip separation control mechanism for any widespread, commercial applications. This is even more important for the Raman spectroscopy or hyperspectral imaging (or in this sense for any other, nonlinear optical applications), as the strength of the Raman emission is proportional to the fourth power of the strength of the electric field of applied light the small changes in the strength of the local near-field enhancement can produce the dramatic fluctuation in the strength of the optical signal.

Besides all of the difficulties in implementation, the advantage of the apertureless mode is in theoretically better resolution than the apertured mode, and in the higher surface-enhanced Raman signal, due to plasmon coupling, which makes this method, theoretically, an ideal molecular Raman nanoprobe, a “holy grail” for life scientists (31,32). The hyperspectral subwavelength Raman imaging is extremely important for further studies in system biology, proteomics, and metabolomics, as it is expected it will for the first time allow identification and spatial positioning of biomolecules within a cell, without the introduction of fluorescence labels. Aside from Raman imaging, this approach is expected to be better for the purpose of fluorescence lifetime imaging (FLIM) (29). The apertureless approach also has significant advantages because the surface plasmon enhancement is driven by the metallic tip and a higher local intensity of the scattered field.

However, to date, there is no commercial instrument available based on the apertureless NSOM principle, because of many technical problems, a significantly smaller photon flux, a low signal/noise ratio, and extreme signal dependance on the tip-sample separation distance. It is expected that with improvements in the control mechanism for keeping the sample-tip separation in the subnanometer range, as well as improvements in laser positioning and further enhancement of the photodetector efficiencies, the apertureless NSOM will become more widespread, with the first commercial instruments expected to be introduced to the market in the near future.

The tip for the apertureless NSOM can be manufactured in three different ways. The simplest way is by electrochemically etching the metallic wire in the same way as for production of the STM tips. However, controlled and optimized shapes, which can provide more efficient field enhancement, can be better achieved by using the free ion bombardment (FIB) techniques for both tip growth and etching.

NSOM Operations

The typical operation of the NSOM consists of positioning the tip above the sample feature of interest, visually using the reflection and transmission video system, respectively. After executing a manual approach procedure, the tip is subsequently placed under PID control and is automatically maintained in the near-field region. The nonoptical, shear-force feedback relies on measuring the voltage generated by a quartz tuning fork onto which the NSOM tip is rigidly mounted, thus avoiding feedback laser. Having a feedback without the feedback laser is of great advantage, as avoidance of that voltage is a direct measure for the oscillation amplitude of the tip-tuning fork assembly, which varies with tip-to-sample distance over a range of ~25 nm.

Unlike conventional AFM cantilever designs, the spring constant of the straight-fiber NSOM tip in the vertical direction is extremely high, thus avoiding damaging snap-to-contact. A feedback algorithm monitors the amplitude of the tuning fork by appropriately adjusting the tip-sample distance. Using this method, the NSOM tip is engaged and maintained within ~5 nm of the surface in the near-field region throughout the NSOM scanning or spectroscopic measurements. Another advantage of using shear-force feedback is in the absence of a feedback laser, which is especially important in the low photon-count applications (e.g., in spectroscopy–hyperspectral imaging and single-molecular studies).

Other methods of maintaining the tip in the near field have not proven nearly as sensitive or reliable as tuning-fork-based shear-force feedback. Some methods originally developed for AFM applications may require actual surface contact and, consequently, possible surface or tip transformation, ultimately resulting in either damage or having to move the tip in and out of the near-field during data acquisition. The other disadvantage of AFM-like feedback is in the great technical difficulties to form a self-actuated, piezo-based AFM hollow tip, thus forcing the use of laser for feedback control. The AFM-like force feedback with pulled fiber tip requires a bent fiber, which is much less mechanically stable than a straight fiber. In addition, the bent fiber has problems associated with circular light paths and higher Raman scattering from glass–fiber substrates, interferences that carry especially negative consequences for near-field spectroscopy, as they can significantly increase the optical noise level.

The NSOM can be used both in air and in liquid. Most of the work to date has been done in air. While it has been demonstrated many times that the method can be successfully used in liquid operation, there are problems associated with the menisc force formed between the probe and liquid surface. For work in air, shear force is the superior method of feedback. However, for work in liquid, the AFM-like noncontact feedback has advantages. With further improvements in the fiber-based probes coating in the near future, it is expected that shear-force-based topographic imaging and feedback will become equal with the AFM-like noncontact-based topographic imaging in liquid.

Near-Field Spectroscopy

For spectroscopic studies, NSOM can be considered as a controlled light collector, with tens-of-nanometer spatial positioning resolution. Thus, many of the standard spectroscopic techniques could be applied, depending on the amount of signal available. For example, it was successfully demonstrated that NSOM can be used for fluorescence and photoluminescence spectroscopy; electroluminescence spectroscopy, time-resolved spectroscopy; polarization studies, and in early stage infrared (IR) and Raman spectroscopies. The latter of the two has the greatest promise in becoming the ultimate molecular nanoprobe. Some of the applications of *in situ* hyperspectral, that is, composition-specific nanoscale studies include chemical identification of observed samples; studies of optical properties of materials at nano-scale levels; detection of phase differences and impurities in materials; protein studies, and many more. Ultimately, it opens the door for a plethora of both fundamental and applied studies in the fields of physics, material science, chemistry, life sciences, and nanotechnology.

Examples of a near-field spectroscopy application are shown in Fig. 7. While Fig. 7a represents the topography image of the PIC dye crystal (33), Fig. 7b is the NSOM fluorescence image of the same area. In order to explore the origin of inhomogeneities, near-field fluorescence spectroscopy, with spectra presented at Fig. 7c, has been done at different points, labeled 1–4, on Fig. 7b. Finally, fluorescence spectra resolved the inhomogenieties of emission sources, pointing to the two different allotropes of crystals having emissions peaking at 645 and 690 nm, respectively.

The near-field signal, which is by default very weak, gets even weaker if we want to do the wavelength-resolved spectroscopic analysis, or full hyperspectral cube imaging, so longer exposition time is necessary to acquire usable spectra. For a near-field spectroscopy system to be successful, it needs to have an extremely stable probe position control, in all three axes, to keep the optimal sample-probe distance, and to keep the probe above the point of interest, for a prolonged time of signal collection.

Figure 8 represents the typical modern commercial NSOM microscopy–spectroscopy setup, in this particular case, an Aurora-3 for spectroscopy made by Veeco Instruments Inc., Santa Barbara, CA (34). In general, such a near-field microscopy–spectroscopy package consists of the NSOM microscope, an objective lens for signal collection, optical filters for elimination of the excitation laser light,

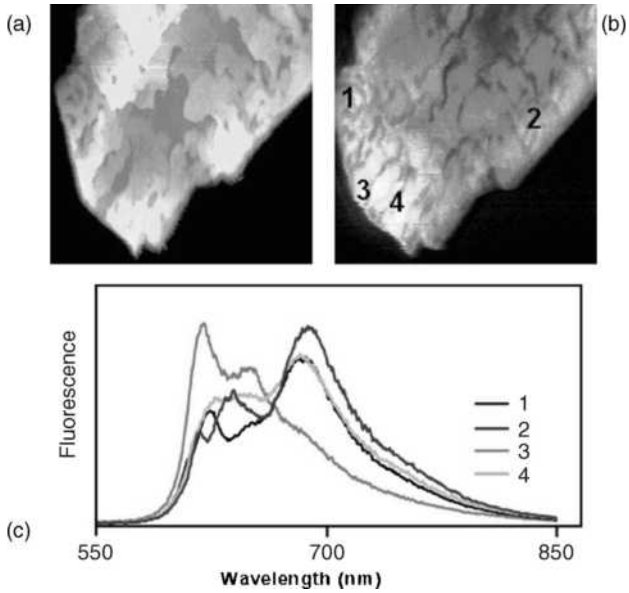


Figure 7. Example of fluorescence-based hyperspectral near-field imaging. (a) Shear-force topography image; (b) NSOM fluorescence image; (c) spatially resolved fluorescence spectra, numbers 1–4 corresponds to the different position on the crystal, demonstrating chemically specific imaging and material inhomogenities at the nanoscale. (Courtesy of David Vanden Bout and Paul Barbara, University of Minnesota, Minneapolis, MI.)

an optical pathway for signal transduction to the detector, and a wavelength dispersive detector. As the signal is generally very weak, the spectrometer needs to have a very sensitive detection system, in the best case, single-photon sensitivity. With today's solid-state detectors technology, the best detector to use is the CCD camera with a larger stack of sensitive pixels coupled to the imaging spectrometer. Depending on applications, either a Peltier cooled camera will be satisfactory (for most bright fluorescence samples) or a liquid-nitrogen-cooled camera for ultimate sensitivity, such as in single-molecular experiments and near-field Raman spectroscopy. Furthermore, interfacing and communication between spectrometer and NSOM scanning probe control software needs to be established.

Today, many spectrometers have semiopen control software, thus allowing triggering of the spectral acquisition with the TTL handshake signal, which can be produced by the NSOM microscope controller. In this way, the microscope controller initiates spectral acquisition and the system can be used with many different commercial spectrometers, allowing customization of the microprobe. Filtering out excitation photons is of extreme importance to increase the signal/noise ratio, and the best filters available today are notch, interferometric filters. Another consideration when designing the NSOM spectroscopy package is that not one formula will fit all of the requirements. Furthermore, virtually any application will require some special design consideration, so building the system

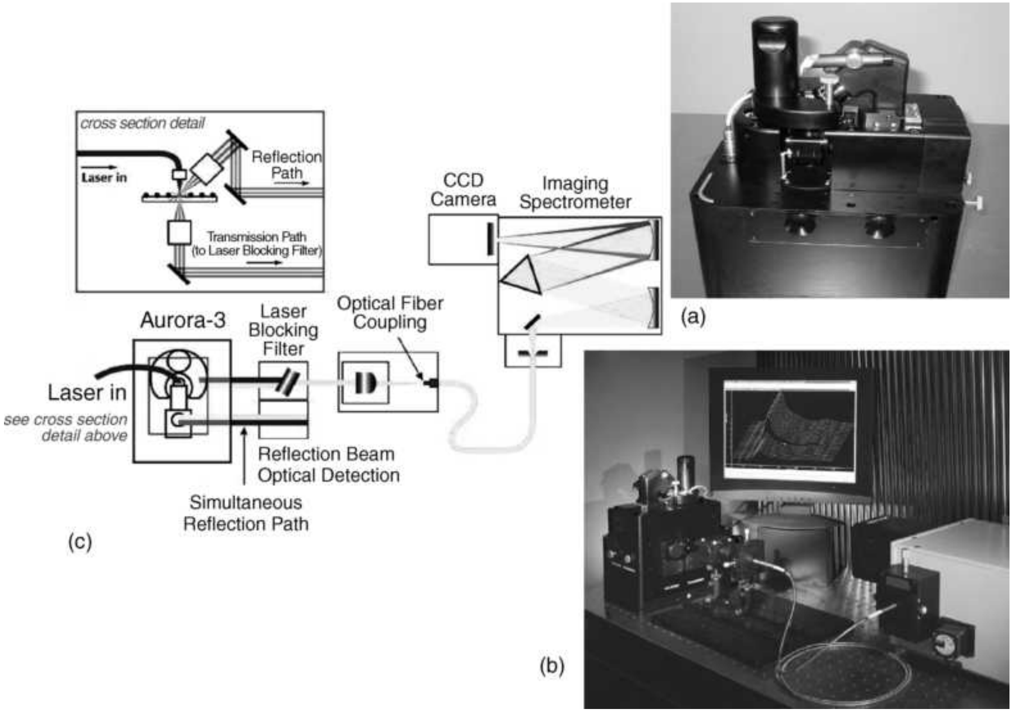


Figure 8. Typical, third generation commercially available NSOM setup, for imaging and spectroscopy: (a) photo of the NSOM head and sample holder; (b) complete system with attached spectrometer; (c) optical path schematics. (Courtesy of Veeco Instruments Inc., Santa Barbara, CA.)

as versatile and modular as possible is of the utmost importance. The ability to incorporate additional standard optical components and easy system reconfiguration should be first in the designer's mind. At the same time, once a system is set up for operation, it should allow for easy, straightforward, simple operation, and robustness, which are sometimes contradictory requirements. The design of the NSOM head, when having the spectroscopy package in mind, must incorporate extremely accurate closed-loop scan linearization. While shear-force feedback keeps superb control of sample-tip distance, the closed-loop X–Y scanning is necessary in order to keep the probe at the selected spatial position during sample collection.

The near-field spectroscopy package is an even more promising breakthrough technology for life sciences than for NSOM imaging alone. For the first time, it will allow qualitative identification of the composition of an observed sample, via optical spectroscopies, at the spatial resolution of scanning probe microscopy. The current commercial systems can distinguish differences in chemical composition of the samples, at the spatial resolution of 50 nm or better. For combined near-field spectroscopy and imaging, the signal may be split into the dual-beam path, wherein one path is used for the imaging detector and another one for spectroscopy. Such dual-beam solution minimizes removal, replacement, and realignment of components when a different mode is desired. The benefit is the ease of use and robust, reliable operation for separate or simultaneous transmission and reflection measurements.

In the example of the Aurora-3 optical path system, the light from both objectives is redirected out the side ports by two front-surface mirrors, which are reflection-coated for optimal visible/near-IR (NIR) operation. The near-field transmission and reflection light is collected in the far field by precisely aligned microscope objectives to provide high quality collimated (parallel) beams, with a nominal 7 mm diameter. This system design allows for NSOM spectroscopic operation with standard one-half in. diameter optics, though for ease of alignment and handling, 1 in. diameter optics is generally recommended. Furthermore, the reflection path is carefully aligned to match the transmission path. Such integrated solutions minimize removal, replacement, and realignment of components when a different mode is desired. The integrated reflection path is always available for use with any sample and never requires removal or realignment in the microscope with normal use. As there are many variations in the spectroscopy setup, it is important for an NSOM system that is intended for spectroscopic use to be capable of utilizing the standard optical element, so users can customize the system using standard optical poles and optical bench mounting systems.

Evaluation

While the first NSOM was invented just 2 years after the AFM, its widespread use has just started to pick up in the last several years. The reasons for this lag are severalfold: the first NSOM images were hard to interpret, and as they were achieved without topography modulation, the contrast in the images was not intuitive; there were no commercial instruments available until the mid-1990s, thus all users needed to build their own systems; the use of the home-built and first commercial instrument and its alignment procedures were cumbersome and complicated for any user who was not skilled in optoelectronics development; and the resolution of the systems depended on each individual sample. However, the field is changing rapidly, and with introduction of the latest, third generation of commercial systems (13–16), such as the Aurora-3 from Veeco Instruments Inc., MultiView 400 from Nanonics Imaging Ltd, Smena from NT-MDT, and AlphaSNOM from WiTec GmbH, the ease of use is comparable with the standard scanning probe microscope and is within the skill set of average life sciences user. Furthermore, with improvement in the serial production of the probes and quality control in recent years, the resolution of the NSOM system is becoming more uniform, and resolution expectations can be met with most samples.

The way to evaluate and measure resolution of the NSOM instrument is by utilizing Fisher's projection masks (35). It is virtually accepted as a standard for evaluating NSOM resolution and quality of image topographic modulation, the latter one by comparing the topographic with the optical image. The Fisher project mask is a regular hexagonal array of metallic spikes (Fig. 9). It is produced by having the monolayer of the monodispersed polymer spheres coated with metallic coating, and the spheres subsequently dissolved with organic solvents. What is left is the regular, closed-packed hexagonal matrix of metallic spikes. In transmission mode, the spike is seen as a dark spot on the optical micrography, while in reflection mode, the spike is a bright spot, and the void space is dark.

The near-field optical imaging obviously provides two great advantages over other types of imaging: its ability to simultaneously acquire topography, in a scanning-force manner, and an optical image. The optical image carries a plethora of different information that can be furthermore extracted. The most important advantage is that there are many different ways to extract direct or indirect information on spatial distribution of chemical composition of the observed sample, even on the single molecular level. Furthermore, at current state-of-the-art commercial NSOM instrumentation, the near-field imaging and spectroscopy can be performed at a resolution at least four times as high as the resolution of the best optical confocal microscopes. The obvious applications in the biomedical field are all of the applications for which the standard confocal and inverted fluorescence microscope is used today, but at the same time, done at much higher resolution (36–40). Examples of life sciences and biomedical applications involve optical ultramicroscopy of cells; optical imaging of cell organelles; imaging and spectroscopy of individual molecules and macromolecules; *in vivo* tracking of molecular events and endocytosis; and high resolution chromosome labeling [i.e., high resolution fluorescence *in situ* hybridization (FISH) (39) applications]. Some of these applications, like molecular tracking, and high resolution FISH are unachievable with other methods.

The molecular tracking applications are based on fluorescence, and in the future, Raman SERS applications will revolutionize our way of understanding how cellular

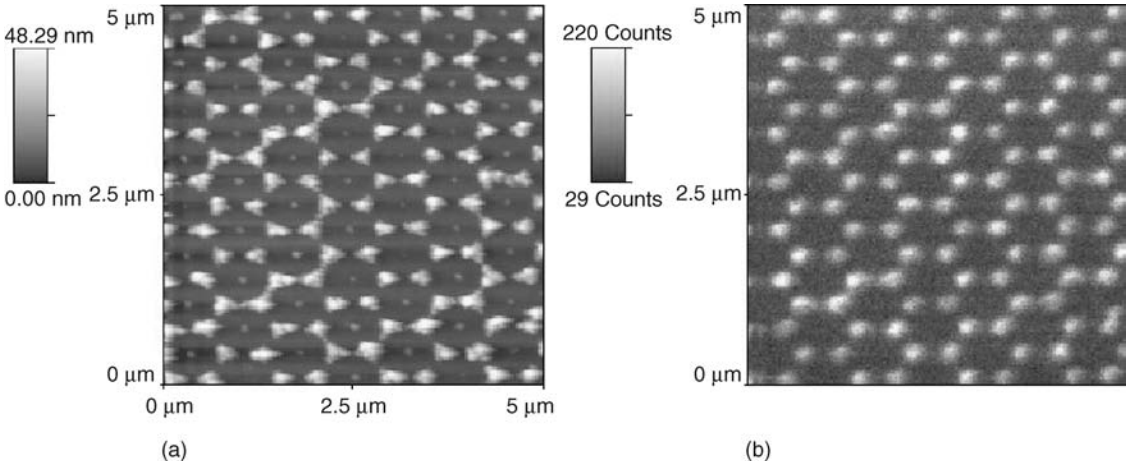


Figure 9. Fisher mask, typical structure for NSOM evaluation and calibration: (a) topographic image produced with shear-force feedback; (b) NSOM transmission mode image of the same area. The images were produced on the Aurora-3 NSOM, (Courtesy of Veeco Instruments Inc., Santa Barbara, CA.)

mechanisms work, and will bring significant contributions to practical pharmacological applications, such as drug candidates, where their target will be able to be imaged, measured, and tracked during action. Additionally, this application will add considerably to the body of knowledge of macromolecular interactions and in the mapping of the interactome of proteins, enzymes, and nucleic acids within the cell. The simplest way will be by expressing differently colored fluorescence proteins, fused with ligand and target, and tracking their spatial positions within the cell, and from the fluorescence resonance energy transfer or fluorescence intermittency, following their interactions. No other method is capable of achieving the resolution necessary to understand molecular machines *in vivo*. Another futuristic application of NSOM is in the ultrahigh density genomics and proteomics array, which theoretically can be packed at the density range higher than the wavelengths.

Some of the more futuristic life science applications will come with the integration of NSOM with mass spectrometry. Zenobi's group (25) demonstrated that the apertured NSOM is capable of doing the nanoablation of structures and thus can feed the mass spectrometer with material ablated with the resolution of a few tens of nanometer. This can dramatically improve the knowledge of spatial locations of proteins, and protein–protein complexes and interactions.

The NSOM-based FISH (38) can have a large impact on the future of molecular *in vitro* diagnostics because it will allow FISH to be applied on the shorter segments of DNA. This will permit many additional applications by painting shorter genes of this simple, chromosome painting technique, which are unachievable with far-field fluorescence or confocal-fluorescence equipment. If designed in a simple and high throughput manner, this may become a standard diagnostic instrument for molecular cytogenetics diagnostics in pathology labs.

It is expected that the next, fourth-generation instrumentation, currently in the design stage, will allow more routine imaging with a level of technical expertise, which is necessary for practical applications comparable with running today's confocal microscope.

Examples of the NSOMs biological applications are presented in Fig. 10a–c. Figure 10a is an example of the protein localization imaging beyond diffraction limit. In this particular image, the fibroblast cells were labeled with green fluorescence protein (GFP). The image on the left represents the shear-force micrography, which corresponds to the topographic image, while the image on the right is the GFP fluorescence image. Spatial distribution of the GFP can easily be observed within the cell at a resolution far exceeding the diffraction limits. This technique can be used to track the protein synthesis and trafficking within the cell if the targeted protein is fused with the fluorescence label, such as GFP or YFP. Another unique NSOM application is in optical characterization of the supramolecular and macromolecular assemblies. Figure 10b represents topographic, shear force (left) and NSOM transmission image (right) of the interband region of a polytene chromosome. In the optical image, the chromatin matter can be distinguished from the DNA based on the optical contrast, which is not possible based on the pure topography. Finally, Fig. 10c represents far-field optical transmission image of slice of the muscle tissue (left) and shear-force topography and near-field optical image, on the right top and bottom, respectively. The near-field imaging reveals the fine structure of the muscular fiber, its cell membrane, myofibrils, and endoplasmatic reticulum structure, in a similar manner as using the transmission electron microscopy.

Besides imaging, the near-field optical microscopy-like setup has promise for use in the nanoscale lithography (41), and for high density data storage (42). Nanoscale lithography will have applications in the preparation of tissue growth matrix and scaffolds, especially for the growth of neurons, while the high voluminous data storage will have a plethora of medical applications for storing ever-growing informational content of both imaging and high throughput diagnostics data.

In conclusion, near-field optical imaging is still a developing technique that shows much promise for biomedical

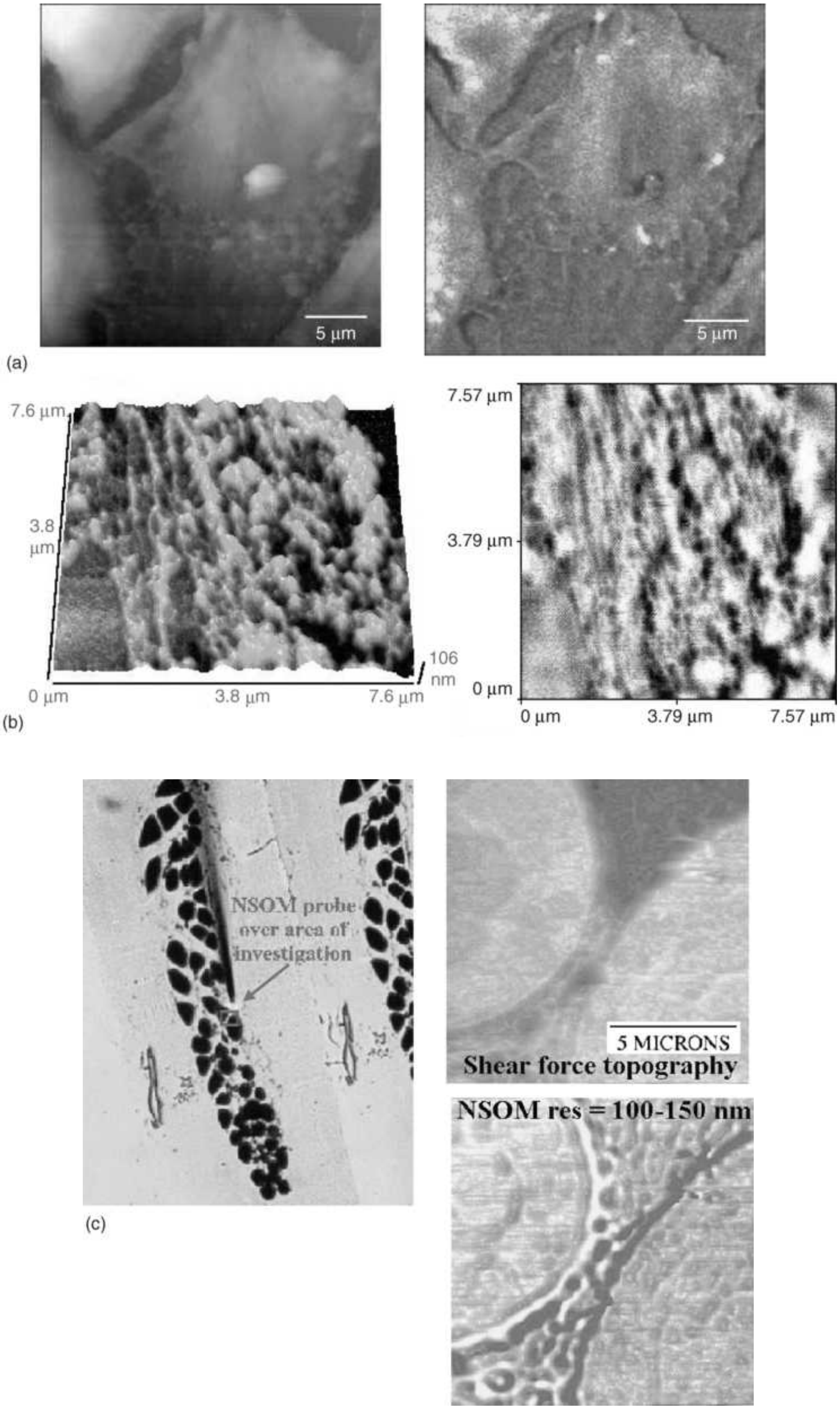


Figure 10. Examples of biomedical and life-sciences applications of NSOM imaging: (a) Shear-force topography and near-field fluorescence from GFP-labeled fibroblast cells. Images of the fibrorblast cells are prepared by growing them directly on glass cover slides and subsequent labeling. They are imaged in air at a scan speed of 1 Hz, no photodegradation was observed throughout the measurement. (Courtesy of Renato Zenobi, ETH, Zurich, © Renato Zenobi and ETH Zurich, Switzerland). (b) Shear force and NSOM image of the interband region of a polytene chromosome. (Courtesy of Sid Ragona and Phil Haydon, Laboratory of Cellular Signaling, Dept. of Zoology and Genetics, Iowa State University, Ames, IA, and Veeco Instruments Inc.). (c) Far-field, differential contrast optical microscopy of the muscle tissue, and details of the muscle cell by shear force and NSOM. (Courtesy of Sid Ragona and Phil Haydon, Laboratory of Cellular Signaling, Department of Zoology and Genetics, Iowa State University, Ames, IA and Veeco Instruments Inc.)

applications. This review presents the state-of-the-art NSOM technology up to the second quarter of 2005. It is certain that by the time of the next edition of this *Encyclopedia*, there will be many other technological and advancements in the field of biomedical applications of near-field optics.

BIBLIOGRAPHY

1. Abbe E. Beitrsge zur theorie des mikroskops und mikroskopichen wahrnehmung. Arch Mikroskop Anat 1873;9:413.
2. Synge EH. A suggested method for extending the microscopic resolution into the ultramicroscopic region. Phil Mag 1928;6: 356–362.
3. Synge EH. An application of piezoelectricity to microscopy. Phil Mag 1932;13:297–300.
4. O'Keefe JA. Resolving power of visible light. J Opt Soc Am 1956;46:359–362.
5. Ash A, Nichols G. Super-resolution aperture scanning microscope. Nature London 1972;237:510–511.
6. Pohl DW, Denk W, Lanz M. Optical stethoscopy—Image recording with resolution lambda/20. Appl Phys Lett 1984;44: 651–653.
7. Lewis A, Isaacson M, Harootunian A, Murray A. Development of 500A spatial-resolution light-microscope. 1. Light is efficiently transmitted through gamma-16 diameter apertures. Ultramicroscopy 1984;13:227–231.
8. Rasmussen A, Deckert V. New dimension in nano-imaging: breaking through the diffraction limit with scanning near-field optical microscopy. Anal Bioanal Chem 2005;381:165–172.
9. Edidin M. Near-field scanning optical microscopy, a siren call to biology. Traffic 2001;2:797–803.
10. Dunn RC. Near-field scanning optical microscopy. Chem Rev 1999;99:2891–2899.
11. Micic M. Near-field scanning optical microscopy and spectroscopy advance. Photonics Spectra 2005;38:124–125.
12. De Serio M, Zenobi R, Deckert V. Looking at the nanoscale: scanning near-field optical microscopy. TRAC-Trends Anal Chem 2003;22:70–77.
13. Product info: Aurora-3 NSOM , Veeco Instruments Inc., Santa Barbara, CA. Available at http://www.veeco.com.
14. Product info: Multiview Series, Nanonics Imaging Ltd., Jerusalem, Israel. Available at http://www.nanonics.co.il.
15. Product info: Alpha SNOM, Witec Wissenschaftliche Instrumente und Technologie. Available at GmbH GmbH,Ulm, Germany, http://www.witec.de.
16. Product info: Solver SNOM, NT-MDT Co, Zelenogorod, Russia. Available at http://www.ntmdt.ru.
17. Paesler MA, Moyer P. Near-Field Optics Theory, Instrumentation and Applications. New York: John Wiley & Sons; 1996.
18. Curie P, Curie J. Developement, par pression de l'electricite polaire dans les cristaux hemiedres a faces inclinees. Comp Ren 1880;91:291–295.
19. Spanner K. Micro Positioning, Nano Positioning, Nano Automation: Solution for Cutting Edge Technologies (product catalog), Karslruhe, Physik Instrumente (PI) GmbH & Co KG. Available at http://www.pi.ws. Accessed 2005.
20. Betzig E, Finn PL, Weiner JS. Combined shear force and near-field scanning optical microscope. Appl Phys Lett 1992;60: 2484–2486.
21. Garcia-Parajo M, Tate T, Chen Y. Gold-coated parabolic tapers for scanning near-field optical microscopy: Fabrication and optimisation. Ultramicroscopy 1995;61:155–163.
22. Pangaribuan T, et al. Reproducible fabrication technique of nanometric tip diameter fiber probe for photon scanning tunneling microscope. Jpn J Appl Phys 1992;31:L1302–L1304.
23. Radojewski R, Grabijec P. Combined SNOM/AFM microscopy with micromachined nanoapertures. Mater Sci–Poland 2003;21:321–332.
24. Takahashi S, Dickson W, Pollard R, Zaytas A. Near-field magneto-optical analysis in reflection mode SNOM. Ultramicroscopy 2004;100(3–4):443–447.
25. Stockle R, et al. Nanoscale atmospheric pressure laser ablation mass spectrometry. Anal Chem 2001;73:139–1402.
26. Chovin A, Garrigue P, Servant L, Sojic N. Electrochemical modulation of remote fluorescence imaging at an ordered opto-electrochemical nanoaperture array. Chemphyschem 2004;5:1125–1132.
27. Inoye Y, Kawata S. Near-field scanning optical microscope with a metallic probe tip. Opt Lett 1994;19:159–161.
28. Keilmann F, Hillenbrand R. Near-field microscopy by elastic light scattering from a tip. Philas Trans R Soc Sci A 2004;362:787–805.
29. Hu DH, et al. Correlated topographic and spectroscopic imaging beyond diffraction limit by atomic force microscopy metallic tip enhanced near-field fluorescence lifetime microscopy. Rev Sci Instr 2003;74:3347–3355.
30. Micic M, Klymyshin N, Suh YD, Lu HP. Finite element method simulation of the field distribution for AFM tip enhanced surface enhanced Raman scanning microscopy. J Phys Chem B 2003;107:1574–1584.
31. Sun WX, Shen ZX. Near-field scanning Raman microscopy using apertureless probes. J Raman Spectrosc 2003;34:668–676.
32. Richards D. Near-field microscopy: Throwing light on the nanoworld Philas Trans R Soc Sci A 2003;361:2843–2857.
33. Vanden Bout DA, Kerimo J, Higgins DA, Barbara PF. Spatially Resolved Spectral Inhomogeneities in Small Molecular Crystals Studied by Near Field Scanning. Opt Microsc J Phys Chem 1996;100:11843–11850.
34. Puestow R. Configuring Aurora-3 for Spectroscopy, application note, Veeco Instruments Inc, Santa Barbara, CA, 2003.
35. Fischer UC, et al. Latex bead projection nanopatterns. Surf Interface Anal 2002;33:75–80.
36. de Lange F, et al. Cell biology beyond the diffraction limit: Near-field scanning optical microscopy. J Cell Sci 2001;114: 4153–4160.
37. Subramaniam V, Kirsch AK, Jovin TM. Cell biological applications of scanning near-field optical microscopy (SNOM). Cell Molec Biol 1998;44:689–700.
38. Lewis A, et al. Near-field scanning optical microscopy in cell biology. Trends Cell Biol 1999;9:70–73.
39. Fukushi D, et al. Scanning near-field optical/atomic force microscopy detection of fluorescence in situ hybridization signals beyond the optical limit. Exp Cell Res 2003;289:237–244.
40. Krishnan RV, Varma R, Mayor S. Fluorescence methods to probe nanometer-scale organization of molecules in living cell membranes. J Fluoresc 2001;11:211–226.
41. Dryakhulshin VF, Klimov AY, Rogov VV, Vostkov NV. Near-field optical lithography method for fabrication of nanodimensional objects. Appl Surf Sci 2005;248:200–203.
42. Ferri V, et al. Near-field optical addressing of luminescent photoswitchable supramolecular system embedded in inert polymer matrices. Nano Lett 2004;4:835–859.

Further Reading

Prasad PN. Nanophotonics. New York: John Wiley & Sons; 2004.

Courion D. Near Field Microscopy and Near Field Optics. London: Imperial College Press; 2003.

Paul DW, Courion D. Near Field Optics. Arc-et Senans: Kulwer Academic Publisher; 1993.

Taatjes DJ, Brooke MT. Cell Imaging Techniques: Methods and Protocolos. Totowa: Humana Press; 2005.

See also MICROSCOPY, CONFOCAL; MICROARRAYS; NANOPARTICLES.

MICROSCOPY, CONFOCAL

NATHAN S. CLAXTON
THOMAS J. FELLERS
MICHAEL W. DAVIDSON
The Florida State University
Tallahassee, Florida

INTRODUCTION

The technique of laser scanning and spinning disk confocal fluorescence microscopy has become an essential tool in biology and the biomedical sciences, as well as in materials science due to attributes that are not readily available using other contrast modes with traditional optical microscopy (1–12). The application of a wide array of new synthetic and naturally occurring fluorochromes has made it possible to identify cells and submicroscopic cellular components with a high degree of specificity amid nonfluorescing material (13). In fact, the confocal microscope is often capable of revealing the presence of a single molecule (14). Through the use of multiply labeled specimens, different probes can simultaneously identify several target molecules simultaneously, both in fixed specimens and living cells and tissues (15). Although both conventional and confocal microscopes cannot provide spatial resolution below the diffraction limit of specific specimen features, the detection of fluorescing molecules below such limits is readily achieved.

The basic concept of confocal microscopy was originally developed by Minsky in the mid-1950s (patented in 1961) when he was a postdoctoral student at Harvard University (16,17). Minsky wanted to image neural networks in unstained preparations of brain tissue and was driven by the desire to image biological events as they occur in living systems. Minsky's invention remained largely unnoticed, due most probably to the lack of intense light sources necessary for imaging and the computer horsepower required to handle large amounts of data. Following Minsky's work, Egger and Petran (18) fabricated a multiple-beam confocal microscope in the late-1960s that utilized a spinning (Nipkow) disk for examining unstained brain sections and ganglion cells. Continuing in this arena, Egger went on to develop the first mechanically scanned confocal laser microscope, and published the first recognizable images of cells in 1973 (19). During the late-1970s and the 1980s, advances in computer and laser technology, coupled to new algorithms for digital manipulation of images, led to a growing interest in confocal microscopy (20).

Fortuitously, shortly after Minsky's patent had expired, practical laser-scanning confocal microscope designs were translated into working instruments by several investigators. Dutch physicist Brakenhoff developed a scanning confocal microscope in 1979 (21), while almost simultaneously, Sheppard contributed to the technique with a theory of image formation (22). Wilson, Amos, and White nurtured the concept and later (during the late-1980s) demonstrated the utility of confocal imaging in the examination of fluorescent biological specimens (20,23). The first commercial instruments appeared in 1987. During the 1990s, advances in optics and electronics afforded more stable and powerful lasers, high efficiency scanning mirror units, high throughput fiber optics, better thin-film dielectric coatings, and detectors having reduced noise characteristics (1). In addition, fluorochromes that were more carefully matched to laser excitation lines were beginning to be synthesized (13). Coupled to the rapidly advancing computer processing speeds, enhanced displays, and large-volume storage technology emerging in the late-1990s, the stage was set for a virtual explosion in the number of applications that could be targeted with laser scanning confocal microscopy.

Modern confocal microscopes can be considered as completely integrated electronic systems where the optical microscope plays a central role in a configuration that consists of one or more electronic detectors, a computer (for image display, processing, output, and storage), and several laser systems combined with wavelength selection devices and a beam scanning assembly. In most cases, integration between the various components is so thorough that the entire confocal microscope is often collectively referred to as a digital or video imaging system capable of producing electronic images (24). These microscopes are now being employed for routine investigations on molecules, cells, and living tissues that were not possible just a few years ago (15).

Confocal microscopy offers several advantages over conventional widefield optical microscopy, including the ability to control depth of field, elimination, or reduction of background information away from the focal plane (that leads to image degradation), and the capability to collect serial optical sections from thick specimens. The basic key to the confocal approach is the use of spatial filtering techniques to eliminate out-of-focus light or glare in specimens whose thickness exceeds the immediate plane of focus. There has been a tremendous explosion in the popularity of confocal microscopy in recent years (1–4,6,7), due in part to the relative ease with which extremely high quality images can be obtained from specimens prepared for conventional fluorescence microscopy, and the growing number of applications in cell biology that rely on imaging, both fixed and living cells and tissues. In fact, confocal technology is proving to be one of the most important advances ever achieved in optical microscopy.

In a conventional widefield optical epi-fluorescence microscope, secondary fluorescence emitted by the specimen often occurs through the excited volume and obscures resolution of features that lie in the objective focal plane (25). The problem is compounded by thicker specimens (>2 μm), which usually exhibit such a high degree of fluorescence emission that most of the fine detail is lost. Confocal microscopy provides only a marginal improvement in both axial (z; parallel to the microscope optical axis) and lateral (x and y; dimensions in the specimen plane) optical resolution, but is able to exclude secondary fluorescence in areas removed from the focal plane from resulting images (26–28). Even though resolution is somewhat enhanced with confocal microscopy over conventional widefield techniques (1), it is still considerably less than that of the transmission electron microscope (TEM). In this regard, confocal microscopy can be considered a bridge between these two classical methodologies.

Illustrated in Fig. 1 are a series of images that compare selected viewfields in traditional widefield and laser

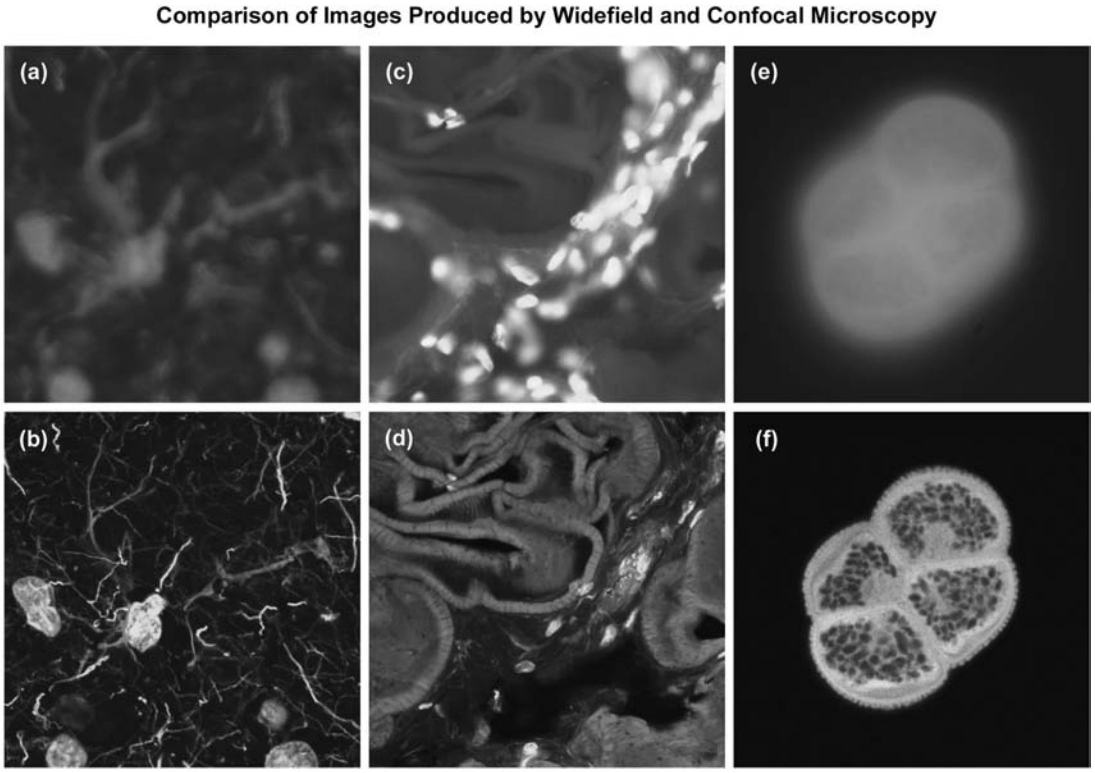

Comparison of Images Produced by Widefield and Confocal Microscopy
(a)
(c)
(e)
(b)
(d)
(f)

Figure 1. Comparison of widefield (upper row) and laser scanning confocal fluorescence microscopy images (lower row). Note the significant amount of signal in the widefield images from fluorescent structures located outside of the focal plane. (a) and (b) Mouse brain hippocampus thick section treated with primary antibodies to glial fibrillary acidic protein (GFAP; red), neurofilaments H (green), and counterstained with Hoechst 33342 (blue) to highlight nuclei. (c) and (d) Thick section of rat smooth muscle stained with phalloidin conjugated to Alexa Fluor 568 (targeting actin; red), wheat germ agglutinin conjugated to Oregon Green 488 (glycoproteins; green), and counterstained with DRAQ5 (nuclei; blue). (e) and (f) Sunflower pollen grain tetrad autofluorescence.

scanning confocal fluorescence microscopy. A thick (16 μm) section of fluorescently stained mouse hippocampus in widefield fluorescence exhibits a large amount of glare from fluorescent structures located above and below the focal plane (Fig. 1a). When imaged with a laser scanning confocal microscope (Fig. 1b), the brain thick section reveals a significant degree of structural detail. Likewise, widefield fluorescence imaging of rat smooth muscle fibers stained with a combination of Alexa Fluor dyes produce blurred images (Fig. 1c) lacking in detail, while the same specimen field (Fig. 1d) reveals a highly striated topography when viewed as an optical section with confocal microscopy. Autofluorescence in a sunflower (*Helianthus annuus*) pollen grain tetrad produces a similar indistinct outline of the basic external morphology (Fig. 1e), but yields no indication of the internal structure in widefield mode. In contrast, a thin optical section of the same grain (Fig. 1f) acquired with confocal techniques displays a dramatic difference between the particle core and the surrounding envelope. Collectively, the image comparisons in Fig. 1 dramatically depict the advantages of achieving very thin optical sections in confocal microscopy. The ability of this technique to eliminate fluorescence emission from regions removed from the focal plane offsets it from traditional forms of fluorescence microscopy.

PRINCIPLES OF CONFOCAL MICROSCOPY

The confocal principle in epi-fluorescence laser scanning microscope is diagrammatically presented in Fig. 2. Coherent light emitted by the laser system (excitation source) passes through a pinhole aperture that is situated in a conjugate plane (confocal) with a scanning point on the specimen and a second pinhole aperture positioned in front of the detector (a photomultiplier tube). As the laser is reflected by a dichromatic mirror, and scanned across the specimen in a defined focal plane, secondary fluorescence emitted from points on the specimen (in the same focal plane) pass back through the dichromatic mirror, and are focused as a confocal point at the detector pinhole aperture.

The significant amount of fluorescence emission that occurs at points above and below the objective focal plane is not confocal with the pinhole (termed out-of-focus light rays in Fig. 2) and forms extended Airy disks in the aperture plane (29). Because only a small fraction of the out-of-focus fluorescence emission is delivered through the pinhole aperture, most of this extraneous light is not detected by the photomultiplier and does not contribute to the resulting image. The dichromatic mirror, barrier filter, and excitation filter perform similar functions to identical components in a widefield epi-fluorescence microscope (30).

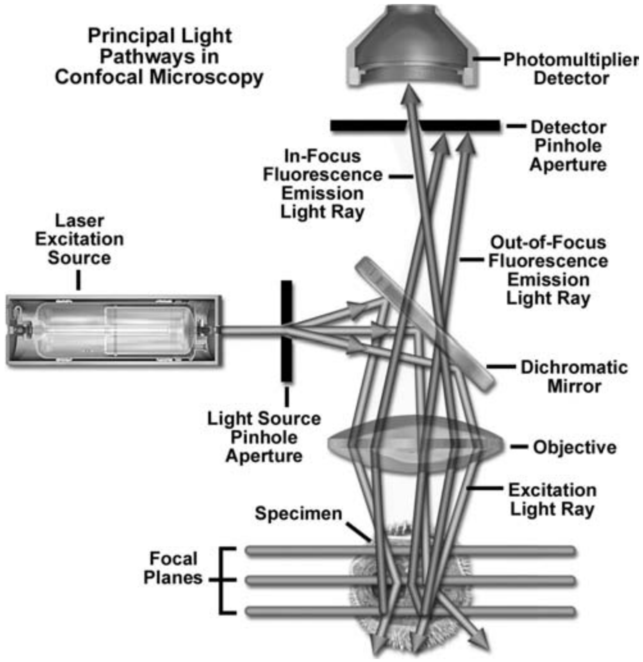

Principal Light Pathways in Confocal Microscopy
Photomultiplier Detector
Detector Pinhole Aperture
In-Focus Fluorescence Emission Light Ray
Laser Excitation Source
Out-of-Focus Fluorescence Emission Light Ray
Dichromatic Mirror
Light Source Pinhole Aperture
Objective
Excitation Light Ray
Specimen
Focal Planes

Figure 2. Schematic diagram of the optical pathway and principal components in a laser scanning confocal microscope.

Refocusing the objective in a confocal microscope shifts the excitation and emission points on a specimen to a new plane that becomes confocal with the pinhole apertures of the light source and detector.

In traditional widefield epi-fluorescence microscopy, the entire specimen is subjected to intense illumination from an incoherent mercury or xenon arc-discharge lamp, and the resulting image of secondary fluorescence emission can be viewed directly in the eyepieces or projected onto the surface of an electronic array detector or traditional film plane. In contrast to this simple concept, the mechanism of image formation in a confocal microscope is fundamentally different (31). As discussed above, the confocal fluorescence microscope consists of multiple laser excitation sources, a scan head with optical and electronic components, electronic detectors (usually photomultipliers), and a computer for acquisition, processing, analysis, and display of images.

The scan head is at the heart of the confocal system and is responsible for rasterizing the excitation scans, as well as collecting the photon signals from the specimen that are required to assemble the final image (1,5–7). A typical scan head contains inputs from the external laser sources, fluorescence filter sets and dichromatic mirrors, a galvanometer-based raster scanning mirror system, variable pinhole apertures for generating the confocal image, and photomultiplier tube detectors tuned for different fluorescence wavelengths. Many modern instruments include diffraction gratings or prisms coupled with slits positioned near the photomultipliers to enable spectral imaging (also referred to as emission fingerprinting) followed by linear unmixing of emission profiles in specimens labeled with combinations of fluorescent proteins or fluorophores having overlapping spectra (32–38). The general arrangement of scan head components is presented in Fig. 3 for a typical commercial unit.

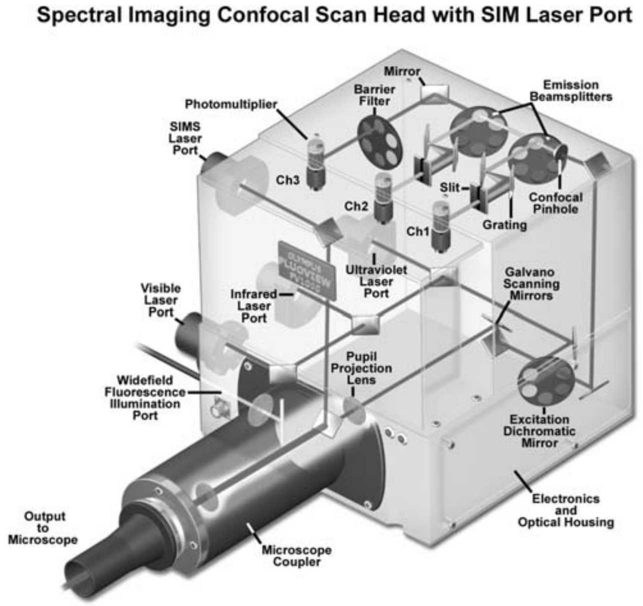


Figure 3. Three-channel spectral imaging laser scanning microscope confocal scan head with SIM scanner laser port. The SIM laser enables simultaneous excitation and imaging of the specimen for photobleaching or photoactivation experiments. Also illustrated are ports for a visible, ultraviolet (UV), and infrared (IR) laser, as well as an arc discharge lamp port for widefield observation.

In epi-illumination scanning confocal microscopy, the laser light source and photomultiplier detectors are both separated from the specimen by the objective, which functions as a well-corrected condenser and objective combination. Internal fluorescence filter components (e.g., the excitation and barrier filters and the dichromatic mirrors) and neutral density filters are contained within the scanning unit (see Fig. 3). Interference and neutral density filters are housed in rotating turrets or sliders that can be inserted into the light path by the operator. The excitation laser beam is connected to the scan unit with a fiber optic coupler followed by a beam expander that enables the thin laser beam wrist to completely fill the objective rear aperture (a critical requirement in confocal microscopy). Expanded laser light that passes through the microscope objective forms an intense diffraction-limited spot that is scanned by the coupled galvanometer mirrors in a raster pattern across the specimen plane (point scanning).

One of the most important components of the scanning unit is the pinhole aperture, which acts as a spatial filter at the conjugate image plane positioned directly in front of the photomultiplier (39). Several apertures of varying diameter are usually contained on a rotating turret that enables the operator to adjust pinhole size (and optical section thickness). Secondary fluorescence collected by the objective is descanned by the same galvanometer mirrors that form the raster pattern, and then passes through a barrier filter before reaching the pinhole aperture (40). The aperture serves to exclude fluorescence signals from out-of-focus features positioned above and below the focal plane, which are instead projected onto the aperture as Airy disks having a diameter much larger than those forming the image. These oversized disks are spread over

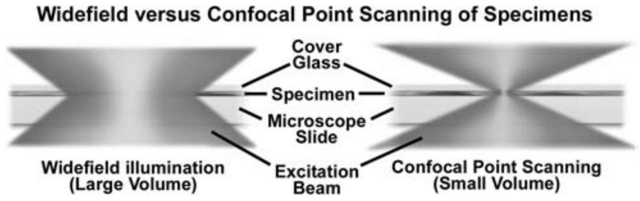

Widefield versus Confocal Point Scanning of Specimens
Cover Glass
Specimen
Microscope Slide
Widefield illumination (Large Volume)
Excitation Beam
Confocal Point Scanning (Small Volume)

Figure 4. Widefield versus confocal microscopy illumination volumes, demonstrating the difference in size between point scanning and widefield excitation light beams.

a comparatively large area so that only a small fraction of light originating in planes away from the focal point passes through the aperture. The pinhole aperture also serves to eliminate much of the stray light passing through the optical system. Coupling of aperture-limited point scanning to a pinhole spatial filter at the conjugate image plane is an essential feature of the confocal microscope.

When contrasting the similarities and differences between widefield and confocal microscopes, it is often useful to compare the character and geometry of specimen illumination utilized for each of the techniques. Traditional widefield epi-fluorescence microscope objectives focus a wide cone of illumination over a large volume of the specimen (41), which is uniformly and simultaneously illuminated (as illustrated in Fig. 4a). A majority of the fluorescence emission directed back toward the microscope is gathered by the objective (depending on the numerical aperture) and projected into the eyepieces or detector. The result is a significant amount of signal due to emitted background light and autofluorescence originating from areas above and below the focal plane, which seriously reduces resolution and image contrast.

The laser illumination source in confocal microscopy is first expanded to fill the objective rear aperture, and then focused by the lens system to a very small spot at the focal plane (Fig. 4b). The size of the illumination point ranges from ~0.25 to 0.8 μm in diameter (depending on the objective numerical aperture) and 0.5 to 1.5 μm deep at the brightest intensity. Confocal spot size is determined by the microscope design, wavelength of incident laser light, objective characteristics, scanning unit settings, and the specimen (41). Figure 4 presents a comparison between the typical illumination cones of a widefield (Fig. 4a) and point scanning confocal (Fig. 4b) microscope at the same numerical aperture. The entire depth of the specimen over a wide area is illuminated by the widefield microscope, while the sample is scanned with a finely focused spot of illumination that is centered in the focal plane in the confocal microscope.

In laser scanning confocal microscopy, the image of an extended specimen is generated by scanning the focused beam across a defined area in a raster pattern controlled by two high speed oscillating mirrors driven with galvanometer motors. One of the mirrors moves the beam from left to right along the x lateral axis, while the other translates the beam in the y direction. After each single scan along the x axis, the beam is rapidly transported back to the starting point and shifted along the y axis to begin a new scan in a process termed flyback (42). During the flyback operation, image information is not collected. In this manner, the area of interest on the specimen in a single focal plane is excited by laser illumination from the scanning unit.

As each scan line passes along the specimen in the lateral focal plane, fluorescence emission is collected by the objective and passed back through the confocal optical system. The speed of the scanning mirrors is very slow relative to the speed of light, so the secondary emission follows a light path along the optical axis that is identical to the original excitation beam. Return of fluorescence emission through the galvanometer mirror system is referred to as descanning (40,42). After leaving the scanning mirrors, the fluorescence emission passes directly through the dichromatic mirror and is focused at the detector pinhole aperture. Unlike the raster scanning pattern of excitation light passing over the specimen, fluorescence emission remains in a steady position at the pinhole aperture, but fluctuates with respect to intensity over time as the illumination spot traverses the specimen producing variations in excitation.

Fluorescence emission that is passed through the pinhole aperture is converted into an analog electrical signal having a continuously varying voltage (corresponding to intensity) by the photomultiplier. The analog signal is periodically sampled and converted into pixels by an analog-to-digital (A/D) converter housed in the scanning unit or the accompanying electronics cabinet. The image information is temporarily stored in an image frame buffer card in the computer and displayed on the monitor. Note that the confocal image of a specimen is reconstructed, point by point, from emission photon signals by the photomultiplier and accompanying electronics, yet never exists as a real image that can be observed through the microscope eyepieces.

LASER SCANNING CONFOCAL MICROSCOPE CONFIGURATION

Basic microscope optical system characteristics have remained fundamentally unchanged for many decades due to engineering restrictions on objective design, the static properties of most specimens, and the fact that resolution is governed by the wavelength of light (1–10). However, fluorescent probes that are employed to add contrast to biological specimens and, and other technologies associated with optical microscopy techniques, have improved significantly. The explosive growth and development of the confocal approach is a direct result of a renaissance in optical microscopy that has been largely fueled by advances in modern optical and electronics technology. Among these are stable multiwavelength laser systems that provide better coverage of the uv, visible, and near-IR spectral regions, improved interference filters (including dichromatic mirrors, barrier, and excitation filters), sensitive low noise wide-band detectors, and far more powerful computers. The latter are now available with relatively low cost memory arrays, image analysis software packages, high resolution video displays, and high quality digital image printers. The flow of information through a modern confocal microscope is presented diagrammatically in Fig. 5 (2).

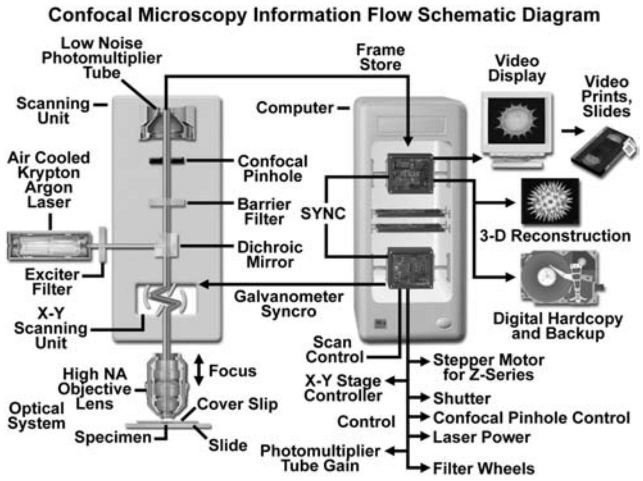


Figure 5. Confocal microscope configuration and information flow schematic diagram.

Although many of these technologies have been developed independently for a variety of specifically targeted applications, they have been incorporated gradually into mainstream commercial confocal microscopy systems. In current microscope systems, classification of designs is based on the technology utilized to scan specimens (7). Scanning can be accomplished either by translating the stage in the *x*, *y*, and *z* directions while the laser illumination spot is held in a fixed position, or the beam itself can be raster-scanned across the specimen. Because three-dimensional (3D) translation of the stage is cumbersome and prone to vibration, most modern instruments employ some type of beam-scanning mechanism.

In modern confocal microscopes, two fundamentally different techniques for beam scanning have been developed. Single-beam scanning, one of the more popular methods employed in a majority of the commercial laser scanning microscopes (43), uses a pair of computer-controlled galvanometer mirrors to scan the specimen in a raster pattern at a rate of approximately one frame per second. Faster scanning rates (to near video speed) can be achieved using acoustooptic devices or oscillating mirrors. In contrast, multiple-beam scanning confocal microscopes are equipped with a spinning Nipkow disk containing an array of pinholes and microlenses (44–46). These instruments often use arc-discharge lamps for illumination instead of lasers to reduce specimen damage and enhance the detection of low fluorescence levels during real-time image collection. Another important feature of the multiple-beam microscopes is their ability to readily capture images with an array detector, such as a charge-coupled device (CCD) camera system (47).

All modern laser scanning confocal microscope designs are centered on a conventional upright or inverted research level optical microscope. However, instead of the standard tungsten–halogen or mercury (xenon) arc-discharge lamp, one or more laser systems are used as a light source to excite fluorophores in the specimen. Image information is gathered point by point with a specialized detector, such as a photomultiplier tube or avalanche photodiode, and then digitized for processing by the host computer, which also controls the scanning mirrors and/or other devices to facilitate the collection and display of images. After a series of images (usually serial optical sections) has been acquired and stored on digital media, analysis can be conducted utilizing numerous image processing software packages available on the host or a secondary computer.

ADVANTAGES AND DISADVANTAGES OF CONFOCAL MICROSCOPY

The primary advantage of laser scanning confocal microscopy is the ability to serially produce thin (0.5–1.5 μm) optical sections through fluorescent specimens that have a thickness ranging up to 50 μm or more (48). The image series is collected by coordinating incremental changes in the microscope fine focus mechanism (using a stepper motor) with sequential image acquisition at each step. Image information is restricted to a well-defined plane, rather than being complicated by signals arising from remote locations in the specimen. Contrast and definition are dramatically improved over widefield techniques due to the reduction in background fluorescence and improved signal to noise (48). Furthermore, optical sectioning eliminates artifacts that occur during physical sectioning and fluorescent staining of tissue specimens for traditional forms of microscopy. The noninvasive confocal optical sectioning technique enables the examination of both living and fixed specimens under a variety of conditions with enhanced clarity.

With most confocal microscopy software packages, optical sections are not restricted to the perpendicular lateral (*x–y*) plane, but can also be collected and displayed in transverse planes (1,5–8,49). Vertical sections in the *x–z* and *y–z* planes (parallel to the microscope optical axis) can be readily generated by most confocal software programs. Thus, the specimen appears as if it had been sectioned in a plane that is perpendicular to the lateral axis. In practice, vertical sections are obtained by combining a series of *x–y* scans taken along the *z* axis with the software, and then projecting a view of fluorescence intensity as it would appear should the microscope hardware have been capable of physically performing a vertical section.

A typical stack of optical sections (often termed a *z series*) through a Lodgepole Pine tree pollen grain revealing internal variations in autofluorescence emission wavelengths is illustrated in Fig. 6. Optical sections were gathered in 1.0 μm steps perpendicular to the *z* axis (microscope optical axis) using a laser combiner featuring an argon ion (488 nm; green fluorescence), a green helium–neon (543 nm; red fluorescence), and a red helium–neon (633 nm; fluorescence pseudocolored blue) laser system. Pollen grains from this and many other species range between 10 and 40 μm in diameter and often yield blurred images in wide-field fluorescence microscopy (see Fig. 1c), which lack information about internal structural details. Although only 12 of the >36 images collected through this series are presented in the figure, they represent individual focal planes separated by a distance of ~3 μm and provide a good indication of the internal grain structure.

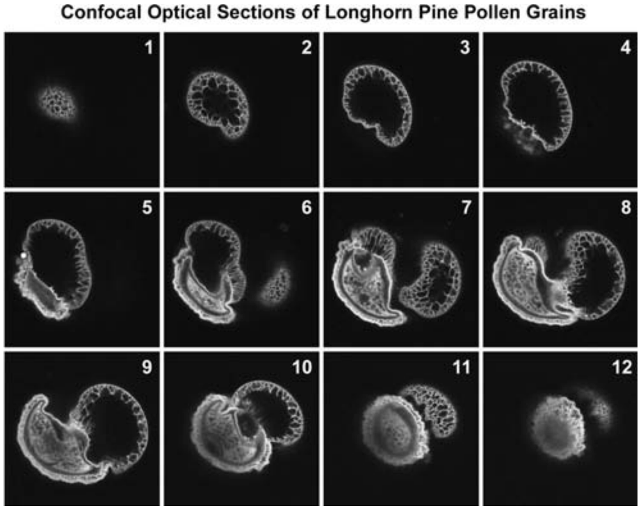


Figure 6. Lodgepole pine (*Pinus contorta*) pollen grain optical sections. Bulk pollen was mounted in CytoSeal 60 and imaged with a 100× oil immersion objective (no zoom) in 1 μm axial steps. Each image in the sequence (1–12) represents the view obtained from steps of 3 μm.

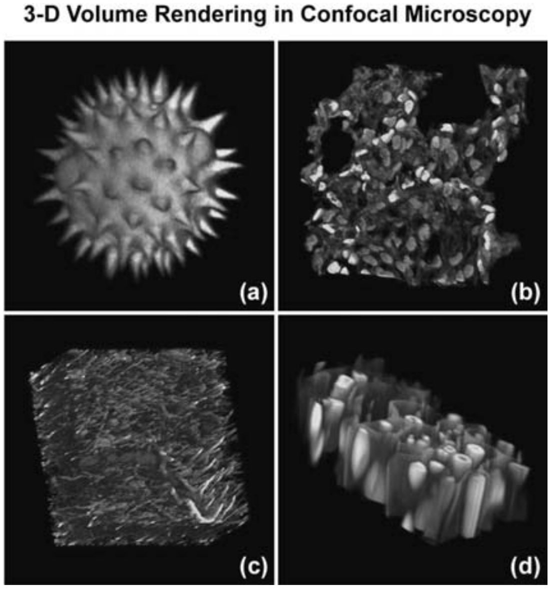


Figure 7. Three-dimensional volume renders from confocal microscopy optical sections. (a) Autofluorescence in a series of sunflower pollen grain optical sections was combined to produce a realistic view of the exterior surface. (b) Mouse lung tissue thick (16 μm) section. (c) Rat brain thick section. These specimens were each labeled with several fluorophores (blue, green, and red fluorescence) and the volume renders were created from a stack of 30–45 optical sections. (d) Autofluorescence in a thin section of fern root.

In specimens more complex than a pollen grain, complex interconnected structural elements can be difficult to discern from a large series of optical sections sequentially acquired through the volume of a specimen with a laser scanning confocal microscope. However, once an adequate series of optical sections has been gathered, it can be further processed into a 3D representation of the specimen using volume-rendering computational techniques (50–53). This approach is now in common use to help elucidate the numerous interrelationships between structure and function of cells and tissues in biological investigations (54). In order to ensure that adequate data is collected to produce a representative volume image, the optical sections should be recorded at the appropriate axial (z step) intervals so that the actual depth of the specimen is reflected in the image.

Most of the software packages accompanying commercial confocal instruments are capable of generating composite and multidimensional views of optical section data acquired from z-series image stacks. The 3D software packages can be employed to create either a single 3D representation of the specimen (Fig. 7) or a video (movie) sequence compiled from different views of the specimen volume. These sequences often mimic the effect of rotation or similar spatial transformation that enhances the appreciation of the specimen's 3D character. In addition, many software packages enable investigators to conduct measurements of length, volume, and depth, and specific parameters of the images, such as opacity, can be interactively altered to reveal internal structures of interest at differing levels within the specimen (54).

Typical 3D representations of several specimens examined by serial optical sectioning are presented in Fig. 7. A series of sunflower pollen grain optical sections was combined to produce a realistic view of the exterior surface (Fig. 7a) as it might appear if being examined by a scanning electron microscope (SEM). The algorithm utilized to construct the 3D model enables the user to rotate the pollen grain through 360° for examination. Similarly, thick sections (16 μm) of lung tissue and rat brain are presented in Fig. 7b and 7c, respectively. These specimens were each labeled with several fluorophores (blue, green, and red fluorescence) and created from a stack of 30–45 optical sections. Autofluorescence in plant tissue was utilized to produce the model illustrated in Fig. 7d of a fern root section.

In many cases, a composite or projection view produced from a series of optical sections provides important information about a 3D specimen than a multidimensional view (54). For example, a fluorescently labeled neuron having numerous thin, extended processes in a tissue section is difficult (if not impossible) to image using wide-field techniques due to out-of-focus blur. Confocal thin sections of the same neuron each reveal portions of several extensions, but these usually appear as fragmented streaks and dots and lack continuity (53). Composite views created by flattening a series of optical sections from the neuron will reveal all of the extended processes in sharp focus with well-defined continuity. Structural and functional analysis of other cell and tissue sections also benefits from composite views as opposed to, or coupled with, 3D volume rendering techniques.

Advances in confocal microscopy have made possible multidimensional views (54) of living cells and tissues that include image information in the x, y, and z dimensions as a function of time and presented in multiple colors (using two or more fluorophores). After volume processing of individual image stacks, the resulting data can be displayed as 3D multicolor video sequences in real time. Note that unlike conventional widefield microscopy, all fluorochromes in multiply labeled specimens appear in register using the confocal microscope. Temporal data can be collected either from time-lapse experiments conducted over extended periods or through real-time image acquisition in smaller frames for short periods of time. The potential for using multidimensional confocal microscopy as a powerful tool in cellular biology is continuing to grow as new laser systems are developed to limit cell damage and computer processing speeds and storage capacity improves.

Additional advantages of scanning confocal microscopy include the ability to adjust magnification electronically by varying the area scanned by the laser without having to change objectives. This feature is termed the zoom factor, and is usually employed to adjust the image spatial resolution by altering the scanning laser sampling period (1,2,8,40,55). Increasing the zoom factor reduces the specimen area scanned and simultaneously reduces the scanning rate. The result is an increased number of samples along a comparable length (55), which increases both the image spatial resolution and display magnification on the host computer monitor. Confocal zoom is typically employed to match digital image resolution (8,40,55) with the optical resolution of the microscope when low numerical aperture and magnification objectives are being used to collect data.

Digitization of the sequential analog image data collected by the confocal microscope photomultiplier (or similar detector) facilitates computer image processing algorithms by transforming the continuous voltage stream into discrete digital increments that correspond to variations in light intensity. In addition to the benefits and speed that accrue from processing digital data, images can be readily prepared for print output or publication. In carefully controlled experiments, quantitative measurements of spatial fluorescence intensity (either statically or as a function of time) can also be obtained from the digital data.

Disadvantages of confocal microscopy are limited primarily to the limited number of excitation wavelengths available with common lasers (referred to as laser lines), which occur over very narrow bands and are expensive to produce in the UV region (56). In contrast, conventional widefield microscopes use mercury- or xenon-based arc-discharge lamps to provide a full range of excitation wavelengths in the UV, visible, and near-IR spectral regions. Another downside is the harmful nature (57) of high intensity laser irradiation to living cells and tissues, an issue that has recently been addressed by multiphoton and Nipkow disk confocal imaging. Finally, the high cost of purchasing and operating multiuser confocal microscope systems (58), which can range up to an order of magnitude higher than comparable widefield microscopes, often limits their implementation in smaller laboratories. This problem can be easily overcome by cost-shared microscope systems that service one or more departments in a core facility. The recent introduction of personal confocal systems has competitively driven down the price of low end confocal microscopes and increased the number of individual users.

CONFOCAL MICROSCOPE LIGHT DETECTORS

In modern widefield fluorescence and laser scanning confocal optical microscopy, the collection and measurement of secondary emission gathered by the objective can be accomplished by several classes of photosensitive detectors (59), including photomultipliers, photodiodes, and solid-state CCDs. In confocal microscopy, fluorescence emission is directed through a pinhole aperture positioned near the image plane to exclude light from fluorescent structures located away from the objective focal plane, thus reducing the amount of light available for image formation, as discussed above. As a result, the exceedingly low light levels most often encountered in confocal microscopy necessitate the use of highly sensitive photon detectors that do not require spatial discrimination, but instead respond very quickly with a high level of sensitivity to a continuous flux of varying light intensity.

Photomultipliers, which contain a photosensitive surface that captures incident photons and produces a stream of photoelectrons to generate an amplified electric charge, are the popular detector choice in many commercial confocal microscopes (59–61). These detectors contain a critical element, termed a photocathode, capable of emitting electrons through the photoelectric effect (the energy of an absorbed photon is transferred to an electron) when exposed to a photon flux. The general anatomy of a photomultiplier consists of a classical vacuum tube in which a glass or quartz window encases the photocathode and a chain of electron multipliers, known as dynodes, followed by an anode to complete the electrical circuit (62). When the photomultiplier is operating, current flowing between the anode and ground (zero potential) is directly proportional to the photoelectron flux generated by the photocathode when it is exposed to incident photon radiation.

In a majority of commercial confocal microscopes, the photomultiplier is located within the scan head or an external housing, and the gain, offset, and dynode voltage are controlled by the computer software interface to the detector power supply and supporting electronics (7). The voltage setting is used to regulate the overall sensitivity of the photomultiplier, and can be adjusted independently of the gain and offset values. The latter two controls are utilized to adjust the image intensity values to ensure that the maximum number of gray levels is included in the output signal of the photomultiplier. Offset adds a positive or negative voltage to the output signal, and should be adjusted so that the lowest signals are near the photomultiplier detection threshold (40). The gain circuit multiplies the output voltage by a constant factor so that the maximum signal values can be stretched to a point just below saturation. In practice, offset should be applied first before adjusting the photomultiplier gain (8,40). After the signal has been processed by the analog-to-digital converter, it is stored in a frame buffer and ultimately displayed on the monitor in a series of gray levels ranging from black (no signal) to white (saturation). Photomultipliers with a dynamic range of 10 or 12 bits are capable of displaying 1024 or 4096 gray levels, respectively. Accompanying image files also have the same number of gray levels. However, the photomultipliers used in a majority of the commercial confocal microscopes have a dynamic range limited to 8 bits or 256 gray levels, which in most cases, is adequate for handling the typical number of photons scanned per pixel (63).

Changes to the photomultiplier gain and offset levels should not be confused with postacquisition image processing to adjust the levels, brightness, or contrast in the final image. Digital image processing techniques can stretch existing pixel values to fill the black-to-white display range, but cannot create new gray levels (40). As a result, when a digital image captured with only 200 out of a possible 4096 gray levels is stretched to fill the histogram (from black to white), the resulting processed image appears grainy. In routine operation of the confocal microscope, the primary goal is to fill as many of the gray levels during image acquisition and not during the processing stages.

The offset control is used to adjust the background level to a position near 0 V (black) by adding a positive or negative voltage to the signal. This ensures that dark features in the image are very close to the black level of the host computer monitor. Offset changes the amplitude of the entire voltage signal, but since it is added to or subtracted from the total signal, it does not alter the voltage differential between the high and low voltage amplitudes in the original signal. For example, with a signal ranging from 4 to 18 V that is modified with an offset setting of 4 V, the resulting signal spans 0–14 V, but the difference remains 14 V.

Figure 8 presents a series of diagrammatic schematics of the unprocessed and adjusted output signal from a photomultiplier and the accompanying images captured with a confocal microscope of a living adherent culture of Indian Muntjac deer skin fibroblast cells treated with MitoTracker Red CMXRos, which localizes specifically in the mitochondria. Figure 8a illustrates the raw confocal image along with the signal from the photomultiplier. After applying a negative offset voltage to the photomultiplier, the signal and image appear in Fig. 8b. Note that as the signal is shifted to lower intensity values, the image becomes darker (upper frame in Fig. 8b). When the gain is adjusted to the full intensity range (Fig. 8c), the image exhibits a significant amount of detail with good contrast and high resolution.

The photomultiplier gain adjustment is utilized to electronically stretch the input signal by multiplying with a constant factor prior to digitization by the analog-to-digital converter (40). The result is a more complete representation of gray level values between black and white, and an increase in apparent dynamic range. If the gain setting is increased beyond the optimal point, the image becomes grainy, but this maneuver is sometimes necessary to capture the maximum number of gray levels present in the image. Advanced confocal microscopy software packages ease the burden of gain and offset adjustment by using a pseudocolor display function to associate pixel values with gray levels on the monitor. For example, the saturated pixels (255) can be displayed in yellow or red, while black-evel pixels (0) are shown in blue or green, with intermediate gray levels displayed in shades of gray representing their true values. When the photomultiplier output is properly adjusted, just a few red (or yellow) and blue (or green) pixels are present in the image, indicating that the full dynamic range of the photomultiplier is being utilized.

Established techniques in the field of enhanced night vision have been applied with dramatic success to photomultipliers designed for confocal microscopy (63,64). Several manufacturers have collaborated to fabricate a head-on photomultiplier containing a specialized prism system that assists in the collection of photons. The prism operates by diverting the incoming photons to a pathway that promotes total internal reflection in the photomultiplier envelope adjacent to the photocathode. This configuration increases the number of potential interactions between the photons and the photocathode, resulting in an increase in quantum efficiency by more than a factor of 2 in the green spectral region, 4 in the red region, and even higher in the IR (59). Increasing the ratio of photoelectrons generated to the number of incoming photons serves to increase the electrical current from the photomultiplier, and to produce a higher sensitivity for the instrument.

Photomultipliers are the ideal photometric detectors for confocal microscopy due to their speed, sensitivity, high signal/noise ratio, and adequate dynamic range (59–61).

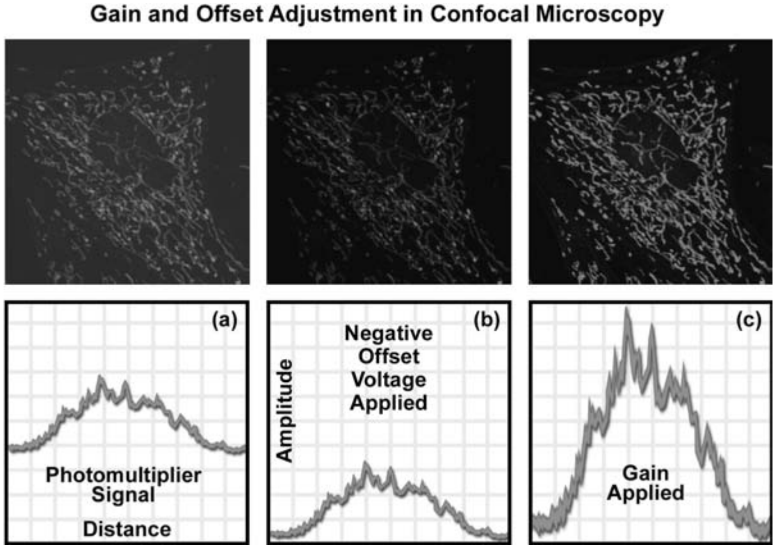

Gain and Offset Adjustment in Confocal Microscopy
(a)
Photomultiplier Signal
Distance
(b)
Negative Offset Voltage Applied
Amplitude
(c)
Gain Applied

Figure 8. Gain and offset control in confocal microscopy photomultiplier detection units. The specimen is a living adherent culture of Indian Muntjac deer skin fibroblast cells treated with MitoTracker Red CMXRos. (a) The raw confocal image (upper frame) along with the signal from the photomultiplier. (b) Signal and confocal image after applying a negative offset voltage to the photomultiplier. (c) Final signal and image after the gain has been adjusted to fill the entire intensity range.

High end confocal microscope systems have several photomultipliers that enable simultaneous imaging of different fluorophores in multiply labeled specimens. Often, an additional photomultiplier is included for imaging the specimen with transmitted light using differential interference or phase-contrast techniques. In general, confocal microscopes contain three photomultipliers for the fluorescence color channels (red, green, and blue; each with a separate pinhole aperture) utilized to discriminate between fluorophores, along with a fourth for transmitted or reflected light imaging. Signals from each channel can be collected simultaneously and the images merged into a single profile that represents the real colors of the stained specimen. If the specimen is also imaged with a brightfield contrast-enhancing technique, such as differential interference contrast (65), the fluorophore distribution in the fluorescence image can be overlaid onto the brightfield image to determine the spatial location of fluorescence emission within the structural domains.

ACOUSTOOPTIC TUNABLE FILTERS IN CONFOCAL MICROSCOPY

The integration of optoelectronic technology into confocal microscopy has provided a significant enhancement in the versatility of spectral control for a wide variety of fluorescence investigations. The acoustooptic tunable filter (AOTF) is an electrooptical device that functions as an electronically tunable excitation filter to simultaneously modulate the intensity and wavelength of multiple laser lines from one or more sources (66). Devices of this type rely on a specialized birefringent crystal whose optical properties vary upon interaction with an acoustic wave. Changes in the acoustic frequency alter the diffraction properties of the crystal, enabling very rapid wavelength tuning, limited only by the acoustic transit time across the crystal.

An acoustooptic tunable filter designed for microscopy typically consists of a tellurium dioxide or quartz anisotropic crystal to which a piezoelectric transducer is bonded (67–70). In response to the application of an oscillating radio frequency (RF) electrical signal, the transducer generates a high frequency vibrational (acoustic) wave that propagates into the crystal. The alternating ultrasonic acoustic wave induces a periodic redistribution of the refractive index through the crystal that acts as a transmission diffraction grating or Bragg diffracter to deviate a portion of incident laser light into a first-order beam, which is utilized in the microscope (or two first-order beams when the incident light is nonpolarized). Changing the frequency of the transducer signal applied to the crystal alters the period of the refractive index variation, and therefore, the wavelength of light that is diffracted. The relative intensity of the diffracted beam is determined by the amplitude (power) of the signal applied to the crystal.

In the traditional fluorescence microscope configuration, including many confocal systems, spectral filtering of both excitation and emission light is accomplished utilizing thin-film interference filters (7). These filters are limiting in several respects. Because each filter has a fixed central wavelength and passband, several filters must be utilized to provide monochromatic illumination for multispectral imaging, as well as to attenuate the beam for intensity control, and the filters are often mechanically interchanged by a rotating turret mechanism. Interference filter turrets and wheels have the disadvantages of limited wavelength selection, vibration, relatively slow switching speed, and potential image shift (70). They are also susceptible to damage and deterioration caused by exposure to heat, humidity, and intense illumination, which changes their spectral characteristics over time. In addition, the utilization of filter wheels for illumination wavelength selection has become progressively more complex and expensive as the number of lasers being employed has increased with current applications.

Rotation of filter wheels and optical block turrets introduces mechanical vibrations into the imaging and illumination system, which consequently requires a time delay for damping of perhaps 50 ms, even if the filter transition itself can be accomplished more quickly. Typical filter change times are considerably slower in practice, however, and range on the order of 0.1–0.5 s. Mechanical imprecision in the rotating mechanism can introduce registration errors when sequentially acquired multicolor images are processed. Furthermore, the fixed spectral characteristics of interference filters do not allow optimization for different fluorophore combinations, nor for adaptation to new fluorescent dyes, limiting the versatility of both the excitation and detection functions of the microscope. Introduction of the AOTF to confocal systems overcomes most of the filter wheel disadvantages by enabling rapid simultaneous electronic tuning and intensity control of multiple laser lines from several lasers.

As applied in laser scanning confocal microscopy, one of the most significant benefits of the AOTF is its capability to replace much more complex and unwieldy filter mechanisms for controlling light transmission, and to apply intensity modulation for wavelength discrimination purposes (67,70). The ability to perform extremely rapid adjustments in the intensity and wavelength of the diffracted beam gives the AOTF unique control capabilities. By varying the illumination intensity at different wavelengths, the response of multiple fluorophores, for example, can be balanced for optimum detection and recording (71). In addition, digital signal processors along with phase and frequency lock-in techniques can be employed to discriminate emission from multiple fluorophores or to extract low level signals from background.

A practical light source configuration scheme utilizing an acoustooptic tunable filter for confocal microscopy is illustrated in Fig. 9. The output of three laser systems (violet diode, argon, and argon–krypton) are combined by dichromatic mirrors and directed through the AOTF, where the first-order diffracted beam (green) is collinear and is launched into a single-mode fiber. The undiffracted laser beams (violet, green, yellow, and red) exit the AOTF at varying angles and are absorbed by a beam stop (not illustrated). The major lines (wavelengths) produced by each laser are indicated (in nm) beneath the hot and cold mirrors. The dichromatic mirror reflects wavelengths < 525 nm and transmits longer wavelengths. Two longer wavelength lines produced by the argon–krypton laser

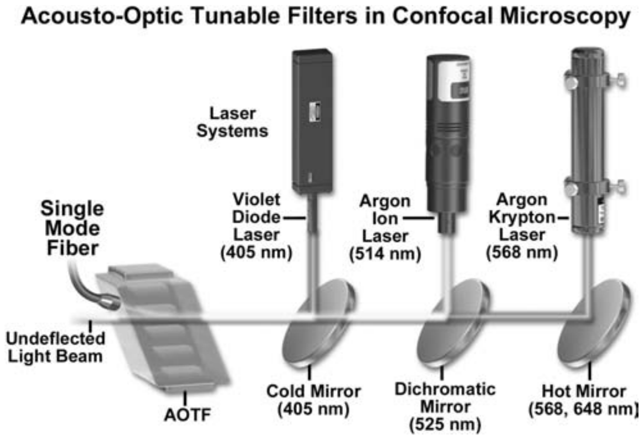

Acousto-Optic Tunable Filters in Confocal Microscopy
Laser Systems
Single Mode Fiber
Violet Diode Laser (405 nm)
Argon Ion Laser (514 nm)
Argon Krypton Laser (568 nm)
Undeflected Light Beam
AOTF
Cold Mirror (405 nm)
Dichromatic Mirror (525 nm)
Hot Mirror (568, 648 nm]

Figure 9. Configuration scheme utilizing an AOTF for laser intensity control and wavelength selection in confocal microscopy.

(568 and 648 nm) are reflected by the hot mirror, while the output of the argon laser (458, 476, 488, and 514 nm) is reflected by the dichromatic mirror and combined with the transmitted light from the argon–krypton laser. Output from the violet diode laser (405 nm) is reflected by the cold mirror and combined with the longer wavelengths from the other two lasers, which are transmitted through the mirror.

Because of the rapid optical response from the AOTF crystal to the acoustic transducer, the acoustooptic interaction is subject to abrupt transitions resembling a rectangular rather than sinusoidal waveform (66). This results in the occurrence of sidelobes in the AOTF passband on either side of the central transmission peak. Under ideal acoustooptic conditions, these sidelobes should be symmetrical about the central peak, with the first lobe having 4.7% of the central peak's intensity. In practice, the sidelobes are commonly asymmetrical and exhibit other deviations from predicted structure caused by variations in the acoustooptic interaction, among other factors. In order to reduce the sidelobes in the passband to insignificant levels, several types of amplitude apodization of the acoustic wave are employed (66,67), including various window functions, which have been found to suppress the highest sidelobe by 30–40 dB. One method that can be used in reduction of sidelobe level with noncollinear AOTFs is to apply spatial apodization by means of weighted excitation of the transducer. In the collinear AOTF, a different approach has been employed, which introduces an acoustic pulse, apodized in time, into the filter crystal.

The effective linear aperture of an AOTF is limited by the acoustic beam height in one dimension (ID) and by the acoustic attenuation across the optical aperture (the acoustic transit distance) in the other dimension (67). The height of the acoustic beam generated within the AOTF crystal is determined by the performance and physical properties of the acoustic transducer. Acoustic attenuation in crystalline materials, such as tellurium dioxide, is proportional to the square of acoustic frequency, and is therefore a more problematic limitation to linear aperture size in the shorter wavelength visible light range, which requires higher RF frequencies for tuning. Near-IR and IR radiation produces less restrictive limitations because of the lower acoustic frequencies associated with diffraction of these longer wavelengths.

The maximum size of an individual acoustic transducer is constrained by performance and power requirements in addition to the geometric limitations of the instrument configuration, and AOTF designers may use an array of transducers bonded to the crystal in order to increase the effective lateral dimensions of the propagating acoustic beam, and to enlarge the area of acoustooptic interaction (66,67,70). The required drive power is one of the most important variables in acoustooptic design, and generally increases with optical aperture and for longer wavelengths. In contrast to acoustic attenuation, which is reduced in the IR spectral range, the higher power required to drive transducers for infrared AOTFs is one of the greatest limitations in these devices. High drive power levels result in heating of the crystal, which can cause thermal drift and instability in the filter performance (66). This is particularly a problem when acoustic power and frequency are being varied rapidly over a large range, and the crystal temperature does not have time to stabilize, producing transient variations in refractive index. If an application requires wavelength and intensity stability and repeatability, the AOTF should be maintained at a constant temperature. One approach taken by equipment manufacturers to minimize this problem is to heat the crystal above ambient temperature, to a level at which it is relatively unaffected by the additional thermal input of the transducer drive power. An alternative solution is to house the AOTF in a thermoelectrically cooled housing that provides precise temperature regulation. Continuing developmental efforts promise to lead to new materials that can provide relatively large apertures combined with effective separation of the filtered and unfiltered beams without use of polarizers, while requiring a fraction of the typical device drive power.

In a noncollinear AOTF, which spatially separates the incident and diffracted light paths, the deflection angle (the angle separating diffracted and undiffracted light beams exiting the crystal) is an additional factor limiting the effective aperture of the device (67). As discussed previously, the deflection angle is greater for crystals having greater birefringence, and determines in part the propagation distance required for adequate separation of the diffracted and undiffracted beams to occur after exiting the crystal. The required distance is increased for larger entrance apertures, and this imposes a practical limit on maximum aperture size because of constraints on the physical dimensions of components that can be incorporated into a microscope system. The angular aperture is related to the total light collecting power of the AOTF, an important factor in imaging systems, although in order to realize the full angular aperture without the use of polarizers in the noncollinear AOTF, its value must be smaller than the deflection angle. Because the acoustooptic tunable filter is not an image-forming component of the microscope system (it is typically employed for source filtering), there is no specific means of evaluating the spatial resolution for this type of device (70). However, the AOTF may restrict the attainable spatial resolution of the imaging system because of its limited linear aperture size and acceptance angle, in the same manner as other optical components. Based on the Rayleigh criterion and the angular and linear apertures of the AOTF, the maximum number of resolvable image elements may be calculated for a given wavelength, utilizing different expressions for the polar and azimuthal planes. Although diffraction limited resolution can be attained in the azimuthal plane, dispersion in the AOTF limits the resolution in the polar plane, and measures must be taken to suppress this factor for optimum performance. The dependence of deflection angle on wavelength can produce one form of dispersion, which is typically negligible when tuning is performed within a relatively narrow bandwidth, but significant in applications involving operation over a broad spectral range. Changes in deflection angle with wavelength can result in image shifts during tuning, producing errors in techniques, such as ratio imaging of fluorophores excited at different wavelengths, and in other multispectral applications. When the image shift obeys a known relationship to wavelength, corrections can be applied through digital processing techniques (1,7). Other effects of dispersion, including reduced angular resolution, may result in image degradation, such as blurring, that requires more elaborate measures to suppress.

SUMMARY OF AOTF BENEFITS IN CONFOCAL MICROSCOPY

Considering the underlying principles of operation and performance factors that relate to the application of AOTFs in imaging systems, a number of virtues from such devices for light control in fluorescence confocal microscopy are apparent. Several benefits of the AOTF combine to greatly enhance the versatility of the latest generation of confocal instruments, and these devices are becoming increasing popular for control of excitation wavelength ranges and intensity. The primary characteristic that facilitates nearly every advantage of the AOTF is its capability to allow the microscopist control of the intensity and/or illumination wavelength on a pixel-by-pixel basis while maintaining a high scan rate (7). This single feature translates into a wide variety of useful analytical microscopy tools, which are even further enhanced in flexibility when laser illumination is employed.

One of the most useful AOTF functions allows the selection of small user-defined specimen areas (commonly termed regions of interest; ROI) that can be illuminated with either greater or lesser intensity, and at different wavelengths, for precise control in photobleaching techniques, excitation ratio studies, resonance energy-transfer investigations, or spectroscopic measurements (see Fig. 10). The illumination intensity can not only be increased in selected regions for controlled photobleaching experiments (71–73), but can be attenuated in desired areas in order to minimize unnecessary photobleaching. When the illumination area is under AOTF control, the laser exposure is restricted to the scanned area by default, and the extremely rapid response of the device can be utilized to provide beam blanking during the flyback interval of the galvanometer scanning mirror cycle, further limiting unnecessary specimen exposure. In practice, the regions of excitation are typically defined by freehand drawing or using tools to produce defined geometrical shapes in an overlay plane on the computer monitor image. Some systems allow any number of specimen areas to be defined for laser exposure, and the laser intensity to be set to different levels for each area, in intensity increments as small as 0.1%. When the

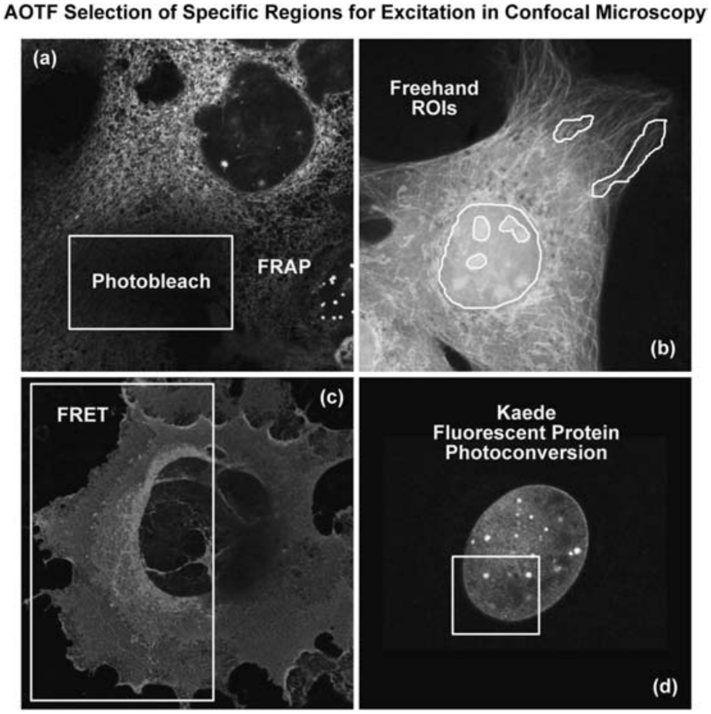

AOTF Selection of Specific Regions for Excitation in Confocal Microscopy
(a)
Photobleach
FRAP
Freehand ROIs
(b)
FRET
(c)
Kaede Fluorescent Protein Photoconversion
(d)

Figure 10. AOTF selection of specific regions for excitation in confocal microscopy. (a) Region of Interest (ROI) selected for fluorescence recovery after photobleaching (FRAP) experiments. (b) Freehand ROIs for selective excitation. (c) ROI for fluorescence resonance energy-transfer (FRET) analysis. (d) ROI for photoactivation and photoconversion of fluorescent proteins.

AOTF is combined with multiple lasers and software that allows time course control of sequential observations, time-lapse experiments can be designed to acquire data from several different areas in a single experiment, which might, for example, be defined to correspond to different cellular organelles.

Figure 10 illustrates several examples of several user-defined ROIs that were created for advanced fluorescence applications in laser scanning confocal microscopy. In each image, the ROI is outlined with a yellow border. The rat kangaroo kidney epithelial cell (PtK2 line) presented in Fig. 10a has a rectangular area in the central portion of the cytoplasm that has been designated for photobleaching experiments. Fluorophores residing in this region can be selectively destroyed by high power laser intensity, and the subsequent recovery of fluorescence back into the photobleached region monitored for determination of diffusion coefficients. Several freehand ROIs are illustrated in Fig. 10b, which can be targets for selective variation of illumination intensities or photobleaching and photoactivation experiments. Fluorescence resonance energy-transfer emission ratios can be readily determined using selected regions in confocal microscopy by observing the effect of bleaching the acceptor fluorescence in these areas (Fig. 10c; African green monkey kidney epithelial cells labeled with Cy3 and Cy5 conjugated to cholera toxin, which localizes in the plasma membrane). The AOTF control of laser excitation in selected regions with confocal microscopy is also useful for investigations of protein diffusion in photoactivation studies (74–76) using fluorescent proteins, as illustrated in Fig. 10d. This image frame presents the fluorescence emission peak of the Kaede protein as it shifts from green to red in HeLa (human cervical carcinoma) cell nuclei using selected illumination (yellow box) with a 405 nanometer violet–blue diode laser.

The rapid intensity and wavelength switching capabilities of the AOTF enable sequential line scanning of multiple laser lines to be performed in which each excitation wavelength can be assigned a different intensity in order to balance the various signal levels for optimum imaging (77). Sequential scanning of individual lines minimizes the time differential between signal acquisitions from the various fluorophores while reducing crossover, which can be a significant problem with simultaneous multiple-wavelength excitation (Fig. 11). The synchronized incorporation of multiple fluorescent probes into living cells has grown into an extremely valuable technique for study of protein–protein interactions, and the dynamics of macromolecular complex assembly. The refinement of techniques for incorporating green fluorescent protein (GFP) and its numerous derivatives into the protein-synthesizing mechanisms of the cell has revolutionized living cell experimentation (78–80). A major challenge in multiple-probe studies using living tissue is the necessity to acquire the complete multispectral data set quickly enough to minimize specimen movement and molecular changes that might distort the true specimen geometry or dynamic sequence of events (32–34). The AOTF provides the speed and versatility to control the wavelength and intensity illuminating multiple specimen regions, and to simultaneously or sequentially scan each at sufficient speed to accurately monitor dynamic cellular processes.

A comparison between the application of AOTFs and neutral density filters (78) to control spectral separation of fluorophore emission spectra in confocal microscopy is presented in Fig. 11. The specimen is a monolayer culture

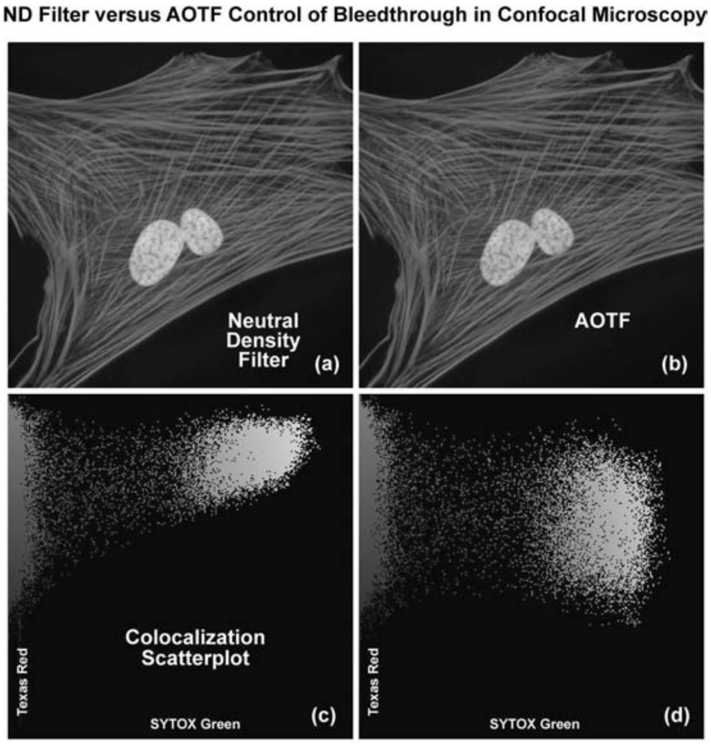

ND Filter versus AOTF Control of Bleedthrough in Confocal Microscopy
Neutral Density Filter
(a)
AOTF
(b)
Colocalization Scatterplot
Texas Red
SYTOX Green
(c)
Texas Red
SYTOX Green
(d)

Figure 11. Fluorophore bleedthrough control with neutral density filters and sequential scanning using AOTF laser modulation. Adherent human lung fibroblast (MRC-5 line) cells were stained with Texas Red conjugated to phalloidin (actin; red) and counterstained with SYTOX green (nuclei; green). (a) Typical cell imaged with neutral density filters. (b) The same cell imaged using sequential line scanning controlled by an AOTF laser combiner. (c) and (d) Colocalization scatterplots derived from the images in (a) and (b), respectively.

of adherent human lung fibroblast (MRC-5 line) cells stained with Texas Red conjugated to phalloidin (targeting the filamentous actin network) and SYTOX Green (staining DNA in the nucleus). A neutral density filter that produces the high excitation signals necessary for both fluorophores leads to a significant amount of bleedthrough of the SYTOX Green emission into the Texas Red channel (Fig. 11a; note the yellow nuclei). The high degree of apparent colocalization between SYTOX Green and Texas Red is clearly illustrated by the scatterplot in Fig. 11b. The two axes in the scatterplot represent the SYTOX Green (abscissa) and the Texas Red (ordinate) channels. In order to balance the excitation power levels necessary to selectively illuminate each fluorophore with greater control of emission intensity, an AOTF was utilized to selectively reduce the SYTOX Green excitation power (Argon-ion laser line at 488 nm). Note the subsequent reduction in bleed-through as manifested by green color in the cellular nuclei in Fig. 11c. The corresponding scatterplot (Fig. 11d) indicates a dramatically reduced level of bleed-through (and apparent colocalization) of SYTOX Green into the Texas Red channel.

The development of the AOTF has provided substantial additional versatility to techniques, such as fluorescence recovery after photobleaching (FRAP; 81,82), fluorescence loss in photobleaching (FLIP; 83), as well as in localized photoactivated fluorescence (uncaging; 84) studies (Fig. 10). The FRAP technique (81,82) was originally conceived to measure diffusion rates of fluorescently tagged proteins in organelles and cell membranes. In the conventional FRAP procedure, a small spot on the specimen is continuously illuminated at a low light flux level and the emitted fluorescence is measured. The illumination level is then increased to a very high level for a brief time to destroy the fluorescent molecules in the illuminated region by rapid bleaching. After the light intensity is returned to the original low level, the fluorescence is monitored to determine the rate at which new unbleached fluorescent molecules diffuse into the depleted region. The technique, as typically employed, has been limited by the fixed geometry of the bleached region, which is often a diffraction-limited spot, and by having to mechanically adjust the illumination intensity (using shutters or galvanometer-driven components). The AOTF not only allows near-instantaneous switching of light intensity, but also can be utilized to selectively bleach randomly specified regions of irregular shape, lines, or specific cellular organelles, and to determine the dynamics of molecular transfer into the region.

By enabling precise control of illuminating beam geometry and rapid switching of wavelength and intensity, the AOTF is a significant enhancement to application of the FLIP technique in measuring the diffusional mobility of certain cellular proteins (83). This technique monitors the loss of fluorescence from continuously illuminated localized regions and the redistribution of fluorophore from distant locations into the sites of depletion. The data obtained can aid in the determination of the dynamic interrelationships between intracellular and intercellular components in living tissue, and such fluorescence loss studies are greatly facilitated by the capabilities of the AOTF in controlling the microscope illumination.

The method of utilizing photoactivated fluorescence has been very useful in studies, such as those examining the role of calcium ion concentration in cellular processes, but has been limited in its sensitivity to localized regional effects in small organelles or in close proximity to cell membranes. Typically, fluorescent species that are inactivated by being bound to a photosensitive species (referred to as being caged) are activated by intense illumination that frees them from the caging compound and allows them to be tracked by the sudden appearance of fluorescence (84). The use of the AOTF has facilitated the refinement of such studies to assess highly localized processes such as calcium ion mobilization near membranes, made possible because of the precise and rapid control of the illumination triggering the activation (uncaging) of the fluorescent molecule of interest.

Because the AOTF functions, without use of moving mechanical components, to electronically control the wavelength and intensity of multiple lasers, great versatility is provided for external control and synchronization of laser illumination with other aspects of microscopy experiments. When the confocal instrument is equipped with a controller module having input and output trigger terminals, laser intensity levels can be continuously monitored and recorded, and the operation of all laser functions can be controlled to coordinate with other experimental specimen measurements, automated microscope stage movements, sequential time-lapse recording, and any number of other operations.

RESOLUTION AND CONTRAST IN CONFOCAL MICROSCOPY

All optical microscopes, including conventional widefield, confocal, and two-photon instruments are limited in the resolution that they can achieve by a series of fundamental physical factors (1,3,5–7,24,85–89). In a perfect optical system, resolution is restricted by the numerical aperture of optical components and by the wavelength of light, both incident (excitation) and detected (emission). The concept of resolution is inseparable from contrast, and is defined as the minimum separation between two points that results in a certain level of contrast between them (24). In a typical fluorescence microscope, contrast is determined by the number of photons collected from the specimen, the dynamic range of the signal, optical aberrations of the imaging system, and the number of picture elements (pixels) per unit area in the final image (66,86–88).

The influence of noise on the image of two closely spaced small objects is further interconnected with the related factors mentioned above, and can readily affect the quality of resulting images (29). While the effects of many instrumental and experimental variables on image contrast, and consequently on resolution, are familiar and rather obvious, the limitation on effective resolution resulting from the division of the image into a finite number of picture elements (pixels) may be unfamiliar to those new to digital microscopy. Because all digital confocal images employing laser scanners and/or camera systems are recorded and processed in terms of measurements made within discrete pixels (66), some discussion of the concepts of sampling theory is required. This is appropriate to the subject of contrast and resolution because it has a direct bearing on the ability to record two closely spaced objects as being distinct.

In addition to the straightforward theoretical aspects of resolution, regardless of how it is defined, the reciprocal relationship between contrast and resolution has practical significance because the matter of interest to most microscopists is not resolution, but visibility. The ability to recognize two closely spaced features as being separate relies on advanced functions of the human visual system to interpret intensity patterns, and is a much more subjective concept than the calculation of resolution values based on diffraction theory (24). Experimental limitations and the properties of the specimen itself, which vary widely, dictate that imaging cannot be performed at the theoretical maximum resolution of the microscope.

The relationship between contrast and resolution with regard to the ability to distinguish two closely spaced specimen features implies that resolution cannot be defined without reference to contrast, and it is this interdependence that has led to considerable ambiguity involving the term resolution and the factors that influence it in microscopy (29). As discussed above, recent advances in fluorescent protein technology have led to an enormous increase in studies of dynamic processes in living cells and tissues (71–76,78–83). Such specimens are optically thick and inhomogeneous, resulting in a far-from-ideal imaging situation in the microscope. Other factors, such as cell viability and sensitivity to thermal damage and photobleaching, place limits on the light intensity and duration of exposure, consequently limiting the attainable resolution. Given that the available timescale may be dictated by these factors and by the necessity to record rapid dynamic events in living cells, it must be accepted that the quality of images will not be as high as those obtained from fixed and stained specimens. The most reasonable resolution goal for imaging in a given experimental situation is that the microscope provides the best resolution possible within the constraints imposed by the experiment.

THE AIRY DISK AND LATERAL RESOLUTION

Imaging a point-like light source in the microscope produces an electromagnetic field in the image plane whose amplitude fluctuations can be regarded as a manifestation of the response of the optical system to the specimen. This field is commonly represented through the amplitude point spread function, and allows evaluation of the optical transfer properties of the combined system components (29,86–88). Although variations in field amplitude are not directly observable, the visible image of the point source formed in the microscope and recorded by its imaging system is the intensity point spread function, which describes the system response in real space. Actual specimens are not point sources, but can be regarded as a superposition of an infinite number of objects having dimensions below the resolution of the system. The properties of the intensity point spread function (PSF; see Fig. 12) in the image plane as well as in the axial direction are major factors in determining the resolution of a microscope (1,24,29,40,85–89).

It is possible to experimentally measure the intensity point spread function in the microscope by recording the image of a subresolution spherical bead as it is scanned through focus (a number of examples may be found in the literature). Because of the technical difficulty posed in direct measurement of the intensity point spread function, calculated point spread functions are commonly utilized to evaluate the resolution performance of different optical systems, as well as the optical-sectioning capabilities of confocal, two-photon, and conventional widefield microscopes. Although the intensity point spread function extends in all three dimensions, with regard to the relationship between resolution and contrast, it is useful to consider only the lateral components of the intensity distribution, with reference to the familiar Airy disk (24).

The intensity distribution of the point-spread function in the plane of focus is described by the rotationally symmetric Airy pattern. Because of the cylindrical symmetry of the microscope lenses, the two lateral components

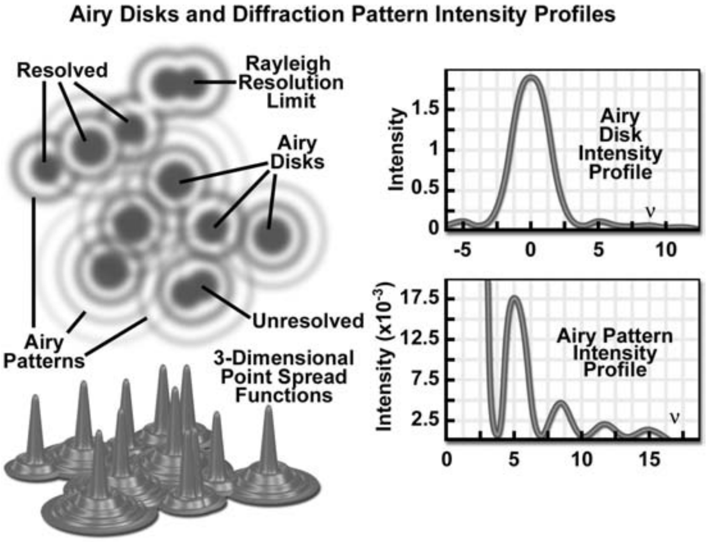

Airy Disks and Diffraction Pattern Intensity Profiles
Resolved
Rayleigh Resolution Limit
Airy Disks
Unresolved
Airy Patterns
3-Dimensional Point Spread Functions
Intensity
1.5
1
0.5
0
-5
0
5
10
Airy Disk Intensity Profile
v
Intensity (x10^{-3})
17.5
12.5
7.5
2.5
0
5
10
15
Airy Pattern Intensity Profile
v

Figure 12. Schematic diagram of an Airy disk diffraction pattern and the corresponding three-dimensional point spread functions for image formation in confocal microscopy. Intensity profiles of a single Airy disk, as well as the first and higher order maxima are illustrated in the graphs.

(x and y) of the Airy pattern are equivalent, and the pattern represents the lateral intensity distribution as a function of distance from the optical axis (24). The lateral distance is normalized by the numerical aperture of the system and the wavelength of light, and therefore is dimensionless. Figure 12 (Airy disk and intensity function) illustrates diagrammatically the formation and characteristics of the Airy disk, the related 3D point spread function, and Airy patterns in the fluorescence microscope. Following the excitation of fluorophores in a point-like specimen region, fluorescence emission occurs in all directions, a small fraction of which is selected and focused by the optical components into an image plane where it forms an Airy disk surrounded by concentric rings of successively decreasing maximum and minimum intensity (the Airy pattern). The Airy pattern intensity distribution is the result of Fraunhofer diffraction of light passing through a circular aperture, and in a perfect optical system exhibits a central intensity maximum and higher order maxima separated by regions of zero intensity (85). The distance of the zero crossings from the optical axis, when the distance is normalized by the numerical aperture and wavelength, occur periodically (Fig. 12). When the intensity on the optical axis is normalized to one (100%), the proportional heights of the first four higher order maxima are 1.7%, 0.4%, 0.2%, and 0.08%, respectively.

A useful approach to the concept of resolution is based on consideration of an image formed by two point-like objects (specimen features), under the assumption that the image-forming process is incoherent, and that the interaction of the separate object images can be described using intensity point spread functions. The resulting image is then composed of the sum of two Airy disks, the characteristics of which depend on the separation distance between the two points (24,86). When sufficiently separated, the intensity change in the area between the objects is the maximum possible, cycling from the peak intensity (at the first point) to zero and returning to the maximum value at the center of the second point. At decreased distance in object space, the intensity distribution functions of the two points, in the image plane, begin to overlap and the resulting image may appear to be that of a single larger or brighter object or feature rather than being recognizable as two objects. If resolution is defined, in general terms, as the minimum separation distance at which the two objects can be sufficiently distinguished, it is obvious that this property is related to the width of the intensity peaks (the point spread function). Microscope resolution is directly related, therefore, to the full-width at half maximum (fwhm) of the instrument's intensity point spread function in the component directions (29,86,87).

Some ambiguity in use of the term resolution results from the variability in defining the degree of separation between features and their point spread functions that is sufficient to allow them to be distinguished as two objects rather than one. In general, minute features of interest in microscopy specimens produce point images that overlap to some extent, displaying two peaks separated by a gap (1,24,29,40,86). The greater the depth of the gap between the peaks, the easier it is to distinguish, or resolve, the two objects. By specifying the depth of the dip in intensity between two overlapping point spread functions, the ambiguity in evaluating resolution can be removed, and a quantitative aspect introduced.

In order to quantify resolution, the concept of contrast is employed, which is defined for two objects of equal intensity as the difference between their maximum intensity and the minimum intensity occurring in the space between them (55,86,89). Because the maximum intensity of the Airy disk is normalized to one, the highest achievable contrast is also one, and occurs only when the spacing between the two objects is relatively large, with sufficient separation to allow the first zero crossing to occur in their combined intensity distribution. At decreased distance, as the two point spread functions begin to overlap, the dip in intensity between the two maxima (and the contrast) is increasingly reduced. The distance at which two peak maxima are no longer discernible, and the contrast becomes zero, is referred to as the contrast cut-off distance (24,40). The variation of contrast with distance allows resolution, in terms of the separation of two points, to be defined as a function of contrast.

The relationship between contrast and separation distance for two point-like objects is referred to as the contrast/distance function or contrast transfer function (31,90). Resolution can be defined as the separation distance at which two objects are imaged with a certain contrast value. It is obvious that when zero contrast exists, the points are not resolved; the so-called Sparrow criterion defines the resolution of an optical system as being equivalent to the contrast cut-off distance (24). It is common, however, to specify that greater contrast is necessary to adequately distinguish two closely spaced points visually, and the well-known Rayleigh criterion (24) for resolution states that two points are resolved when the first minimum (zero crossing) of one Airy disk is aligned with the central maximum of the second Airy disk. Under optimum imaging conditions, the Rayleigh criterion separation distance corresponds to a contrast value of 26.4%. Although any contrast value >0 can be specified in defining resolution, the 26% contrast of the Rayleigh criterion is considered reasonable in typical fluorescence microscopy applications, and is the basis for the common expression defining lateral resolution according to the following equation (24), in which the point separation (r) in the image plane is the distance between the central maximum and the first minimum in the Airy disk:

$$r_{\text{lateral}} = 1.22\,\lambda/(2\cdot\text{NA}) = 0.6\,\lambda/\text{NA}$$

where λ is the emitted light wavelength and NA is the numerical aperture of the objective.

Resolution in the microscope is directly related to the fwhm dimensions of the microscope's point spread function, and it is common to measure this value experimentally in order to avoid the difficulty in attempting to identify intensity maxima in the Airy disk. Measurements of resolution utilizing the fwhm values of the point spread function are somewhat smaller than those calculated employing the Rayleigh criterion. Furthermore, in confocal fluorescence configurations, single-point illumination scanning and single-point detection are employed, so that only the fluorophores in the shared volume of the illumination and detection point spread functions are able to be detected. The intensity point spread function in the confocal case is, therefore, the product of the independent illumination intensity and detection intensity point spread functions. For confocal fluorescence, the lateral (and axial) extent of the point spread function is reduced by ~30% compared to that in the wide-field microscope. Because of the narrower intensity point spread function, the separation of points required to produce acceptable contrast in the confocal microscope (29,31) is reduced to a distance approximated by

$$r_{\text{lateral}} = 0.4\lambda/\text{NA}$$

If the illumination and fluorescence emission wavelengths are approximately the same, the confocal fluorescence microscope Airy disk size is the square of the wide-field microscope Airy disk. Consequently, the contrast cut-off distance is reduced in the confocal arrangement, and equivalent contrast can be achieved at a shorter distance compared to the widefield illumination configuration. Regardless of the instrument configuration, the lateral resolution displays a proportional relationship to wavelength, and is inversely proportional to the objective lens numerical aperture.

As noted previously, lateral resolution is of primary interest in discussing resolution and contrast, although the axial extent of the microscope intensity point spread function is similarly reduced in the confocal arrangement as compared to the widefield fluorescence configuration (86,89). Reasonable contrast between point-like objects lying on the optical axis occurs when they are separated by the distance between the central maximum and the first minimum of the axial point spread function component.

Figure 13 presents the axial intensity distributions (89) for a typical widefield (Fig. 13a) and confocal (Fig. 13b) fluorescence microscope. Note the dramatic reduction in intensity of the wings in the confocal distribution as a function of distance from the central maximum.

A variety of equations are presented in the literature that pertains to different models for calculating axial resolution for various microscope configurations. The ones most applicable to fluorescence emission are similar in form to the expressions evaluating depth of field, and demonstrate that axial resolution is proportional to the wavelength, and refractive index of the specimen medium, and inversely proportional to the square of the numerical aperture. Consequently, the NA of the microscope objective has a much greater effect on axial resolution than does the emission wavelength. One equation (89) commonly used to describe axial resolution for the confocal configuration is given below, with η representing the index of refraction, and the other variables as specified previously:

$$r_{\text{axial}} = 1.4\ \lambda{\cdot}\eta/\text{NA}^2$$

Although the confocal microscope configuration exhibits only a modest improvement in measured axial resolution over that of the widefield microscope, the true advantage of the confocal approach is in the optical sectioning capability in thick specimens, which results in a dramatic improvement in effective axial resolution over conventional techniques. The optical sectioning properties of the confocal microscope result from the characteristics of the integrated intensity point spread function, which has a maximum in the focal plane when evaluated as a function of depth. The equivalent integral of intensity point spread function for the conventional widefield microscope is constant as a function of depth, producing no optical sectioning capabilities.

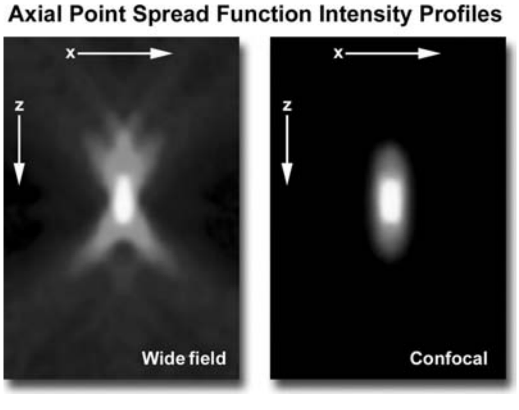


Figure 13. Comparison of axial (*x–z*) point spread functions for widefield (left) and confocal (right) microscopy.

FLUOROPHORES FOR CONFOCAL MICROSCOPY

Biological laser scanning confocal microscopy relies heavily on fluorescence as an imaging mode, primarily due to the high degree of sensitivity afforded by the technique coupled with the ability to specifically target structural components and dynamic processes in chemically fixed as well as living cells and tissues. Many fluorescent probes are constructed around synthetic aromatic organic chemicals designed to bind with a biological macromolecule (e.g., a protein or nucleic acid) or to localize within a specific structural region, such as the cytoskeleton, mitochondria, Golgi apparatus, endoplasmic reticulum, and nucleus (90). Other probes are employed to monitor dynamic processes and localized environmental variables, including concentrations of inorganic metallic ions, pH, reactive oxygen species, and membrane potential (91). Fluorescent dyes are also useful in monitoring cellular integrity (live versus dead and apoptosis), endocytosis, exocytosis, membrane fluidity, protein trafficking, signal transduction, and enzymatic activity (92). In addition, fluorescent probes have been widely applied to genetic mapping and chromosome analysis in the field of molecular genetics.

The history of synthetic fluorescent probes dates back over a century to the late-1800s when many of the cornerstone dyes for modern histology were developed. Among these were pararosaniline, methyl violet, malachite green, safranin O, methylene blue, and numerous azo (nitrogen) dyes, such as Bismarck brown (93). Although these dyes were highly colored and capable of absorbing selected bands of visible light, most were only weakly fluorescent and would not be useful for the fluorescence microscopes that would be developed several decades later. However, several synthetic dye classes synthesized during this period, based on the xanthene and acridine heterocyclic ring systems, proved to be highly fluorescent and provided a foundation for the development of modern synthetic fluorescent probes. Most notable among these early fluorescent dyes were the substituted xanthenes, fluorescein and rhodamine B, and the biaminated acridine derivative, acridine orange.

Fluorochromes were introduced to fluorescence microscopy in the early twentieth century as vital stains for bacteria, protozoa, and trypanosomes, but did not see widespread use until the 1920s when fluorescence microscopy was first used to study dye binding in fixed tissues and living cells (7,93). However, it was not until the early 1940s that Coons developed a technique for labeling antibodies with fluorescent dyes, thus giving birth to the field of immunofluorescence (94). Over the past 60 years, advances in immunology and molecular biology have produced a wide spectrum of secondary antibodies and provided insight into the molecular design of fluorescent probes targeted at specific regions within macromolecular complexes.

Fluorescent probe technology and cell biology were dramatically altered by the discovery of the GFP from jellyfish and the development of mutant spectral variants, which have opened the door to noninvasive fluorescence multicolor investigations of subcellular protein localization, intermolecular interactions, and trafficking using living cell cultures (79,80,95). More recently, the development of nanometer-sized fluorescent semiconductor quantum dots has provided a new avenue for research in confocal and widefield fluorescence microscopy (96). Despite the numerous advances made in fluorescent dye synthesis during the past few decades, there is very little solid evidence about molecular design rules for developing new fluorochromes, particularly with regard to matching absorption spectra to available confocal laser excitation wavelengths. As a result, the number of fluorophores that have found widespread use in confocal microscopy is a limited subset of the many thousands that have been discovered.

BASIC CHARACTERISTICS OF FLUOROPHORES

Fluorophores are catalogued and described according to their absorption and fluorescence properties, including the spectral profiles, wavelengths of maximum absorbance and emission, and the fluorescence intensity of the emitted light (92). One of the most useful quantitative parameters for characterizing absorption spectra is the molar extinction coefficient (denoted with the Greek symbole, see Fig. 14a), which is a direct measure of the ability of a molecule to absorb light. The extinction coefficient is useful for converting units of absorbance into units of molar concentration, and is determined by measuring the absorbance at a reference wavelength (usually the maximum, characteristic of the absorbing species) for a molar concentration in a defined optical path length. The quantum yield of a fluorochrome or fluorophore represents a quantitative measure of fluorescence emission efficiency, and is expressed as the ratio of the number of photons emitted to the number of photons absorbed. In other words, the quantum yield represents the probability that a given excited fluorochrome will produce an emitted (fluorescence) photon. Quantum yields typically range between a value of 0 and 1 and fluorescent molecules commonly employed as probes in microscopy have quantum yields ranging from very low (0.05 or less) to almost unity. In general, a high quantum yield is desirable in most imaging applications. The quantum yield of a given fluorophore varies, sometimes to large extremes, with environmental factors, such as metallic ion concentration, pH, and solvent polarity (92).

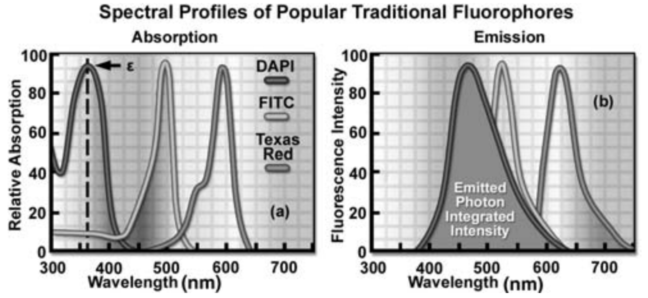


Figure 14. Fluorescent spectral profiles, plotted as normalized absorption or emission as a function of wavelength, for popular synthetic fluorophores emitting in the blue, green, and red regions of the visible spectrum. Each profile is identified with a colored bullet in (a), which illustrates excitation spectra. (b) The emission spectra for the fluorophores according to the legend in (a).

In most cases, the molar extinction coefficient for photon absorption is quantitatively measured and expressed at a specific wavelength, whereas the quantum efficiency is an assessment of the total integrated photon emission over the entire spectral band of the fluorophore (Fig. 14b). As opposed to traditional arc-discharge lamps used with the shortest range (10–20 nm) bandpass interference filters in wide-field fluorescence microscopy, the laser systems used for fluorophore excitation in scanning confocal microscopy restrict excitation to specific laser spectral lines that encompass only a few nanometers (1,7). The fluorescence emission spectrum for both techniques, however, is controlled by similar bandpass or longpass filters that can cover tens to hundreds of nanometers (7). Below saturation levels, fluorescence intensity is proportional to the product of the molar extinction coefficient and the quantum yield of the fluorophore, a relationship that can be utilized to judge the effectiveness of emission as a function of excitation wavelength(s). These parameters display an approximate 20-fold range in variation for the popular fluorophores commonly employed for investigations in confocal microscopy with quantum yields ranging from 0.05 to 1.0, and extinction coefficients ranging from 10,000 to 0.25 million ($L \cdot mol^{-1}$). In general, the absorption spectrum of a fluorophore is far less dependent on environmental conditions than the fluorescence emission characteristics (spectral wavelength profile and quantum yield; 92).

Fluorophores chosen for confocal applications must exhibit a brightness level and signal persistence sufficient for the instrument to obtain image data that does not suffer from excessive photobleaching artifacts and low signal/noise ratios. In widefield fluorescence microscopy, excitation illumination levels are easily controlled with neutral

Table 1. Laser and Arc-Discharge Spectral Lines in Widefield and Confocal Microscopy

Laser Type	Ultraviolet	Violet	Blue	Green	Yellow	Orange	Red
Argon-ion	351, 364		457, 477, 488	514			
Blue diode		405, 440					
Diode-pumped solid state	355	430, 442	457, 473	532	561		
Helium–cadmium	322, 354	442					
Krypton–argon			488		568		647
Green helium–neon				543			
Yellow helium–neon					594		
Orange helium–neon						612	
Red helium-neon							633
Red diode							635, 650
Mercury arc	365	405, 436	546		579		
Xenon arc		467					

density filters (40), and the intensity can be reduced (coupled with longer emission signal collection periods) to avoid saturation and curtail irreversible loss of fluorescence. Excitation conditions in confocal microscopy are several orders of magnitude more severe, however, and restrictions imposed by characteristics of the fluorophores and efficiency of the microscope optical system become the dominating factor in determining excitation rate and emission collection strategies (1,7,92).

Because of the narrow and wavelength-restricted laser spectral lines employed to excite fluorophores in confocal microscopy (Table 1), fluorescence emission intensity can be seriously restricted due to poor overlap of the excitation wavelengths with the fluorophore absorption band. In addition, the confocal pinhole aperture, which is critical in obtaining thin optical sections at high signal/noise ratios, is responsible for a 25–50% loss of emission intensity, regardless of how much effort has been expended on fine-tuning and alignment of the microscope optical system (7). Photomultiplier tubes are the most common detectors in confocal microscopy, but suffer from a quantum efficiency that varies as a function of wavelength (especially in the red and IR regions), further contributing to a wavelength-dependent loss of signal across the emission spectrum (59–62). Collectively, the light losses in confocal microscopy can result in a reduction of intensity exceeding 50 times of the level typically observed in widefield fluorescence instruments. It should be clear from the preceding argument that fluorophore selection is one of the most critical aspects of confocal microscopy, and instrumental efficiency must be carefully considered, as well, in order to produce high quality images.

In confocal microscopy, irradiation of the fluorophores with a focused laser beam at high power densities increases the emission intensity up to the point of dye saturation, a condition whose parameters are dictated by the excited state lifetime (97). In the excited state, fluorophores are unable to absorb another incident photon until they emit a lower energy photon through the fluorescence process. When the rate of fluorophore excitation exceeds the rate of emission decay, the molecules become saturated and the ground state population decreases. As a result, a majority of the laser energy passes through the specimen undiminished and does not contribute to fluorophore excitation. Balancing fluorophore saturation with laser light intensity levels is, therefore, a critical condition for achieving the optimal signal/noise ratio in confocal experiments (1,7,92,97). The number of fluorescent probes currently available for confocal microscopy runs in the hundreds (90,93), with many dyes having absorption maxima closely associated with common laser spectral lines (90). An exact match between a particular laser line and the absorption maximum of a specific probe is not always possible, but the excitation efficiency of lines near the maximum is usually sufficient to produce a level of fluorescence emission that can be readily detected.

Instrumentally, fluorescence emission collection can be optimized by careful selection of objectives, detector aperture dimensions, dichromatic and barrier filters, as well as maintaining the optical train in precise alignment (63). In most cases, low magnification objectives with a high numerical aperture should be chosen for the most demanding imaging conditions because light collection intensity increases as the fourth power of the numerical aperture, but only decreases as the square of the magnification. However, the most important limitations in light collection efficiency in confocal microscopy arise from restrictions imposed by the physical properties of the fluorophores themselves. As previously discussed, fluorescent probe development is limited by a lack of knowledge of the specific molecular properties responsible for producing optimum fluorescence characteristics, and the design rules are insufficiently understood to be helpful as a guide to the development of more efficient fluorophores. The current success in development of new fluorescent probes capable of satisfactory performance in confocal microscopy is a testament to the progress made through use of empirical data and assumptions about molecular structure extrapolated from the properties of existing dyes, many of which were first synthesized over a hundred years ago.

TRADITIONAL FLUORESCENT DYES

The choice of fluorescent probes for confocal microscopy must address the specific capabilities of the instrument to excite and detect fluorescence emission in the wavelength regions made available by the laser systems and detectors. Although the current lasers used in confocal microscopy (Table 1) produce discrete lines in the UV, visible, and near-IR portions of the spectrum, the location of these spectral lines does not always coincide with absorption maxima of popular fluorophores. In fact, it is not necessary for the laser spectral line to correspond exactly with the fluorophore wavelength of maximum absorption, but the intensity of fluorescence emission is regulated by the fluorophore extinction coefficient at the excitation wavelength (as discussed above). The most popular lasers for confocal microscopy are air-cooled argon and krypton–argon ion lasers, the new blue diode lasers, and a variety of helium–neon systems (7,40). Collectively, these lasers are capable of providing excitation at 10–12 specific wavelengths between 400 and 650 nm.

Many of the classical fluorescent probes that have been successfully utilized for many years in widefield fluorescence (92,93), including fluorescein isothiocyanate, Lissamine rhodamine, and Texas red, are also useful in confocal microscopy. Fluorescein is one of the most popular fluorochromes ever designed, and has enjoyed extensive application in immunofluorescence labeling. This xanthene dye has an absorption maximum at 495 nm, which coincides quite well with the 488 nm (blue) spectral line produced by argon-ion and krypton–argon lasers, as well as the 436 and 467 principal lines of the mercury and xenon arc-discharge lamps, respectively. In addition, the quantum yield of fluorescein is very high and a significant amount of information has been gathered on the characteristics of this dye with respect to the physical and chemical properties (98). On the negative side, the fluorescence emission intensity of fluorescein is heavily influenced by environmental factors (e.g., pH), and the relatively broad emission spectrum often overlaps with those of other fluorophores in dual and triple labeling experiments (92,98,99).

Tetramethyl rhodamine (TMR) and the isothiocyanate derivative (TRITC) are frequently employed in multiple labeling investigations in widefield microscopy due to their efficient excitation by the 546 nm spectral line from mercury arc-discharge lamps. The fluorochromes, which have significant emission spectral overlap with fluorescein, can be excited very effectively by the 543 nm line from helium–neon lasers, but not by the 514 or 568 nanometer lines from argon-ion and krypton–argon lasers (99). When using krypton-based laser systems, Lissamine rhodamine is a far better choice in this fluorochrome class due to the absorption maximum at 575 nm and its spectral separation from fluorescein. Also, the fluorescence emission intensity of rhodamine derivatives is not as dependent upon strict environmental conditions as that of fluorescein.

Several of the acridine dyes, first isolated in the nineteenth century, are useful as fluorescent probes in confocal microscopy (93). The most widely utilized, acridine orange, consists of the basic acridine nucleus with dimethylamino substituents located at the 3 and 6 positions of the trinuclear ring system. In physiological pH ranges, the molecule is protonated at the heterocyclic nitrogen and exists predominantly as a cationic species in solution. Acridine orange binds strongly to DNA by intercalation of the acridine nucleus between successive base pairs, and exhibits green fluorescence with a maximum wavelength of 530 nm (92,93,100). The probe also binds strongly to ribonucleic acid (RNA) or single-stranded deoxyribonucleic acid (DNA), but has a longer wavelength fluorescence maximum (∼ 640 nm; red) when bound to these macromolecules. In living cells, acridine orange diffuses across the cell membrane (by virtue of the association constant for protonation) and accumulates in the lysosomes and other acidic vesicles. Similar to most acridines and related polynuclear nitrogen heterocycles, acridine orange has a relatively broad absorption spectrum, which enables the probe to be used with several wavelengths from the argon-ion laser.

Another popular traditional probe that is useful in confocal microscopy is the phenanthridine derivative, propidium iodide, first synthesized as an antitrypanosomal agent along with the closely related ethidium bromide). Propidium iodide binds to DNA in a manner similar to the acridines (via intercalation) to produce orange-red fluorescence centered at 617 nm (101,102). The positively charged fluorophore also has a high affinity for double-stranded RNA. Propidium has an absorption maximum at 536 nm, and can be excited by the 488 or 514-nm spectral lines of an argon-ion (or krypton–argon) laser, or the 543 nm line from a green helium–neon laser. The dye is often employed as a counterstain to highlight cell nuclei during double or triple labeling of multiple intracellular structures. Environmental factors can affect the fluorescence spectrum of propidium, especially when the dye is used with mounting media containing glycerol. The structurally similar ethidium bromide, which also binds to DNA by intercalation (102), produces more background staining, and is therefore not as effective as propidium.

The DNA and chromatin can also be stained with dyes that bind externally to the double helix. The most popular fluorochromes in this category are 4′,6-diamidino-2-phenylindole (DAPI) and the bis (benzimide) Hoechst dyes that are designated by the numbers 33258, 33342, and 34580 (103–106). These probes are quite water soluble and bind externally to AT-rich base pair clusters in the minor groove of double-stranded DNA with a dramatic increase in fluorescence intensity. Both dye classes can be stimulated by the 351 nm spectral line of high power argon-ion lasers or the 354 nm line from a helium–cadmium laser. Similar to the acridines and phenanthridines, these fluorescent probes are popular choices as a nuclear counterstain for use in multicolor fluorescent labeling protocols. The vivid blue fluorescence emission produces dramatic contrast when coupled to green, yellow, and red probes in adjacent cellular structures.

ALEXA FLUOR DYES

The dramatic advances in modern fluorophore technology are exemplified by the Alexa Fluor dyes (90,107,108) introduced by Molecular Probes (Alexa Fluor is a registered trademark of Molecular Probes). These sulfonated rhodamine derivatives exhibit higher quantum yields for more intense fluorescence emission than spectrally similar probes, and have several additional improved features, including enhanced photostability, absorption spectra matched to common laser lines, pH insensitivity, and a high degree of water solubility. In fact, the resistance to photobleaching of Alexa Fluor dyes is so dramatic (108) that even when subjected to irradiation by high intensity laser sources, fluorescence intensity remains stable for relatively long periods of time in the absence of antifade reagents. This feature enables the water soluble Alexa Fluor probes to be readily utilized for both live-cell and tissue section investigations, as well as in traditional fixed preparations.

Alexa Fluor dyes are available in a broad range of fluorescence excitation and emission wavelength maxima, ranging from the UV and deep blue to the near-IR regions (90). Alphanumeric names of the individual dyes are associated with the specific excitation laser or arc-discharge lamp spectral lines for which the probes are intended. For example, Alexa Fluor 488 is designed for excitation by the blue 488 nm line of the argon or krypton–argon ion lasers, while Alexa Fluor 568 is matched to the 568 nm spectral line of the krypton–argon laser. Several of the Alexa Fluor dyes are specifically designed for excitation by either the blue diode laser (405 nm), the orange/yellow helium–neon laser (594 nm), or the red helium–neon laser (633 nm). Other Alexa Fluor dyes are intended for excitation with traditional mercury arc-discharge lamps in the visible (Alexa Fluor 546) or UV (Alexa Fluor 350, also useful with high power argon-ion lasers), and solid-state red diode lasers (Alexa Fluor 680). Because of the large number of available excitation and emission wavelengths in the Alexa Fluor series, multiple labeling experiments can often be conducted exclusively with these dyes.

Alexa Fluor dyes are commercially available as reactive intermediates in the form of maleimides, succinimidyl esters, and hydrazides, as well as prepared cytoskeletal probes (conjugated to phalloidin, G-actin, and rabbit skeletal muscle actin) and conjugates to lectin, dextrin, streptavidin, avidin, biocytin, and a wide variety of secondary antibodies (90). In the latter forms, the Alexa Fluor fluorophores provide a broad palette of tools for investigations in immunocytochemistry, neuroscience, and cellular biology. The family of probes has also been extended into a series of dyes having overlapping fluorescence emission maxima targeted at sophisticated confocal microscopy detection systems with spectral imaging and linear unmixing capabilities. For example, Alexa Fluor 488, Alexa Fluor 500, and Alexa Fluor 514 are visually similar in color with bright green fluorescence, but have spectrally distinct emission profiles. In addition, the three fluorochromes can be excited with the 488 or 514 nm spectral line from an argon-ion laser and are easily detected with traditional fluorescein filter combinations. In multispectral (*x–y–l*; referred to as a lambda stack) confocal imaging experiments, optical separation software can be employed to differentiate between the similar signals (32–35). The overlapping emission spectra of Alexa Fluor 488, 500, and 514 are segregated into separate channels and differentiated using pseudocolor techniques when the three fluorophores are simultaneously combined in a triple label investigation.

CYANINE DYES

The popular family of cyanine dyes, Cy2, Cy3, Cy5, Cy7, and their derivatives, are based on the partially saturated indole nitrogen heterocyclic nucleus with two aromatic units being connected via a polyalkene bridge of varying carbon number (92,109). These probes exhibit fluorescence excitation and emission profiles that are similar to many of the traditional dyes, such as fluorescein and tetramethylrhodamine, but with enhanced water solubility, photostability, and higher quantum yields. Most of the cyanine dyes are more environmentally stable than their traditional counterparts, rendering their fluorescence emission intensity less sensitive to pH and organic mounting media. In a manner similar to the Alexa Fluors, the excitation wavelengths of the Cy series of synthetic dyes are tuned specifically for use with common laser and arc-discharge sources, and the fluorescence emission can be detected with traditional filter combinations.

Marketed by a number of distributors, the cyanine dyes are readily available as reactive dyes or fluorophores coupled to a wide variety of secondary antibodies, dextrin, streptavidin, and eggwhite avidin (110). The cyanine dyes generally have broader absorption spectral regions than members of the Alexa Fluor family, making them somewhat more versatile in the choice of laser excitation sources for confocal microscopy (7). For example, using the 547 nm spectral line from an argon-ion laser, Cy2 is about twice as efficient in fluorescence emission as Alexa Fluor 488. In an analogous manner, the 514 nm argon-ion laser line excites Cy3 with a much higher efficiency than Alexa Fluor 546, a spectrally similar probe. Emission profiles of the cyanine dyes are comparable in spectral width to the Alexa Fluor series.

Included in the cyanine dye series are the long-wavelength Cy5 derivatives, which are excited in the red region (650 nm) and emit in the far-red (680 nm) wavelengths. The Cy5 fluorophore is very efficiently excited by the 647 nm spectral line of the krypton–argon laser, the 633 nm line of the red helium–neon laser, or the 650 nm line of the red diode laser, providing versatility in laser choice. Because the emission spectral profile is significantly removed from traditional fluorophores excited by UV and blue illumination, Cy5 is often utilized as a third fluorophore in triple labeling experiments. However, similar to other probes with fluorescence emission in the far-red spectral region, Cy5 is not visible to the human eye and can only be detected electronically (using a specialized CCD camera system or photomultiplier). Therefore, the probe is seldom used in conventional widefield fluorescence experiments.

FLUORESCENT ENVIRONMENTAL PROBES

Fluorophores designed to probe the internal environment of living cells have been widely examined by a number of investigators, and many hundreds have been developed to monitor such effects as localized concentrations of alkali and alkaline earth metals, heavy metals (employed biochemically as enzyme cofactors), inorganic ions, thiols, and sulfides, nitrite, as well as pH, solvent polarity, and membrane potential (7,90–93,111,112). Originally, the experiments in this arena were focused on changes in the wavelength and/or intensity of absorption and emission spectra exhibited by fluorophores upon binding calcium ions in order to measure intracellular flux densities. These probes bind to the target ion with a high degree of specificity to produce the measured response and are often referred to as spectrally sensitive indicators. Ionic concentration changes are determined by the application of optical ratio signal analysis to monitor the association equilibrium between the ion and its host. The concentration values derived from this technique are largely independent of instrumental variations and probe concentration fluctuations due to photobleaching, loading parameters, and cell retention. In the past few years, a number of new agents have been developed that bind specific ions or respond with measurable features to other environmental conditions (7,90).

Calcium is a metabolically important ion that plays a vital role in cellular response to many forms of external stimuli (113). Because transient fluctuations in calcium ion concentration are typically involved when cells undergo a response, fluorophores must be designed to measure not only localized concentrations within segregated compartments, but should also produce quantitative changes when flux density waves progress throughout the entire cytoplasm. Many of the synthetic molecules designed to measure calcium levels are based on the nonfluorescent chelation agents EGTA and BAPTA, which have been used for years to sequester calcium ions in buffer solutions (7,114,115). Two of the most common calcium probes are the ratiometric indicators fura-2 and indo-1, but these fluorophores are not particularly useful in confocal microscopy (7,116). The dyes are excited by UV light and exhibit a shift in the excitation or emission spectrum with the formation of isosbestic points when binding calcium. However, the optical aberrations associated with UV imaging, limited specimen penetration depths, and the expense of ultraviolet lasers have limited the utility of these probes in confocal microscopy.

Fluorophores that respond in the visible range to calcium ion fluxes are, unfortunately, not ratiometric indicators and do not exhibit a wavelength shift (typical of fura-2 and indo-1) upon binding, although they do undergo an increase or decrease in fluorescence intensity. The best example is fluo-3, a complex xanthene derivative, which undergoes a dramatic increase in fluorescence emission at 525 nm (green) when excited by the 488 nm spectral line of an argon-ion or krypton–argon laser (7,117). Because isosbestic points are not present to assure the absence of concentration fluctuations, it is impossible to determine whether spectral changes are due to complex formation or a variation in concentration with fluo-3 and similar fluorophores.

To overcome the problems associated with using visible light probes lacking wavelength shifts (and isosbestic points), several of these dyes are often utilized in combination for calcium measurements in confocal microscopy (118). Fura red, a multinuclear imidazole and benzofuran heterocycle, exhibits a decrease in fluorescence at 650 nm when binding calcium. A ratiometric response to calcium ion fluxes can be obtained when a mixture of fluo-3 and fura red is excited at 488 nm and fluorescence is measured at the emission maxima (525 and 650 nm, respectively) of the two probes. Because the emission intensity of fluo-3 increases monotonically while that of fura red simultaneously decreases, an isosbestic point is obtained when the dye concentrations are constant within the localized area being investigated. Another benefit of using these probes together is the ability to measure fluorescence intensity fluctuations with a standard FITC/Texas red interference filter combination.

Quantitative measurements of ions other than calcium, such as magnesium, sodium, potassium, and zinc, are conducted in an analogous manner using similar fluorophores (7,90,92). One of the most popular probes for magnesium, mag-fura-2 (structurally similar to fura red), is also excited in the ultraviolet range and presents the same problems in confocal microscopy as fura-2 and indo-1. Fluorophores excited in the visible light region are becoming available for the analysis of many monovalent and divalent cations that exist at varying concentrations in the cellular matrix. Several synthetic organic probes have also been developed for monitoring the concentration of simple and complex anions.

Important fluorescence monitors for intracellular pH include a pyrene derivative known as HPTS or pyranine, the fluorescein derivative, BCECF, and another substituted xanthene termed carboxy SNARF-1 (90,119–122). Because many common fluorophores are sensitive to pH in the surrounding medium, changes in fluorescence intensity that are often attributed to biological interactions may actually occur as a result of protonation. In the physiological pH range (pH 6.8–7.4), the probes mentioned above are useful for dual-wavelength ratiometric measurements and differ only in dye loading parameters. Simultaneous measurements of calcium ion concentration and pH can often be accomplished by combining a pH indicator, such as SNARF-1, with a calcium ion indicator (e.g., fura-2). Other probes have been developed for pH measurements in subcellular compartments, such as the lysosomes, as described below.

ORGANELLE PROBES

Fluorophores targeted at specific intracellular organelles, such as the mitochondria, lysosomes, Golgi apparatus, and endoplasmic reticulum, are useful for monitoring a variety of biological processes in living cells using confocal microscopy (7,90,92). In general, organelle probes consist of a fluorochrome nucleus attached to a target-specific moiety that assists in localizing the fluorophore through covalent, electrostatic, hydrophobic, or similar types of bonds. Many of the fluorescent probes designed for selecting organelles are able to permeate or sequester within the cell membrane (and therefore, are useful in living cells), while others must be installed using monoclonal antibodies with traditional immunocytochemistry techniques. In living cells, organelle probes are useful for investigating transport, respiration, mitosis, apoptosis, protein degradation, acidic compartments, and membrane phenomena. Cell impermeant fluorophore applications include nuclear functions, cytoskeletal structure, organelle detection, and probes for membrane integrity. In many cases, living cells that have been labeled with permeant probes can subsequently be fixed and counterstained with additional fluorophores in multicolor labeling experiments.

Mitochondrial probes are among the most useful fluorophores for investigating cellular respiration and are often employed along with other dyes in multiple labeling investigations. The traditional probes, rhodamine 123 and tetramethylrosamine, are rapidly lost when cells are fixed and have largely been supplanted by newer, more specific, fluorophores developed by Molecular Probes (90,123,124). These include the popular MitoTracker and MitoFluor series of structurally diverse xanthene, benzoxazole, indole, and benzimidazole heterocycles that are available in a variety of excitation and emission spectral profiles. The mechanism of action varies for each of the probes in this series, ranging from covalent attachment to oxidation within respiring mitochondrial membranes.

MitoTracker dyes are retained quite well after cell fixation in formaldehyde and can often withstand lipophilic permeabilizing agents (123). In contrast, the MitoFluor probes are designed specifically for actively respiring cells and are not suitable for fixation and counterstaining procedures (90). Another popular mitochondrial probe, entitled JC-1, is useful as an indicator of membrane potential and in multiple staining experiments with fixed cells (125). This carbocyanine dye exhibits green fluorescence at low concentrations, but can undergo intramolecular association within active mitochondria to produce a shift in emission to longer (red) wavelengths. The change in emission wavelength is useful in determining the ratio of active to nonactive mitochondria in living cells.

In general, weakly basic amines that are able to pass through membranes are the ideal candidates for investigating biosynthesis and pathogenesis in lysosomes (90–92,112). Traditional lysosomal probes include the nonspecific phenazine and acridine derivatives neutral red and acridine orange, which are accumulated in the acidic vesicles upon being protonated (92,93). Fluorescently labeled latex beads and macromolecules, such as dextran, can also be accumulated in lysosomes by endocytosis for a variety of experiments. However, the most useful tools for investigating lysosomal properties with confocal microscopy are the LysoTracker and LysoSensor dyes developed by Molecular Probes (90,92,126). These structurally diverse agents contain heterocyclic and aliphatic nitrogen moieties that modulate transport of the dyes into the lysosomes of living cells for both short- and long-term studies. The LysoTracker probes, which are available in a variety of excitation and emission wavelengths (91), have high selectivity for acidic organelles and are capable of labeling cells at nanomolar concentrations. Several of the dyes are retained quite well after fixing and permeabilization of cells. In contrast, the LysoSensor fluorophores are designed for studying dynamic aspects of lysosome function in living cells. Fluorescence intensity dramatically increases in the LysoSensor series upon protonation, making these dyes useful as pH indicators (91). A variety of Golgi apparatus specific monoclonal antibodies have also been developed for use in immunocytochemistry assays (90,127–129).

Proteins and lipids are sorted and processed in the Golgi apparatus, which is typically stained with fluorescent derivatives of ceramides and sphingolipids (130). These agents are highly lipophilic, and are therefore useful as markers for the study of lipid transport and metabolism in live cells. Several of the most useful fluorophores for Golgi apparatus contain the complex heterocyclic BODIPY nucleus developed by Molecular Probes (90,92,131). When coupled to sphingolipids, the BODIPY fluorophore is highly selective and exhibits a tolerance for photobleaching that is far superior to many other dyes. In addition, the emission spectrum is dependent upon concentration (shifting from green to red at higher concentrations), making the probes useful for locating and identifying intracellular structures that accumulate large quantities of lipids. During live-cell experiments, fluorescent lipid probes can undergo metabolism to derivatives that may bind to other subcellular features, a factor that can often complicate the analysis of experimental data.

The most popular traditional probes for endoplasmic reticulum fluorescence analysis are the carbocyanine and xanthene dyes, DiOC (6) and several rhodamine derivatives, respectively (90,92). These dyes must be used with caution, however, because they can also accumulate in the mitochondria, Golgi apparatus, and other intracellular lipophilic regions. Newer, more photostable, probes have been developed for selective staining of the endoplasmic reticulum by several manufacturers. In particular, oxazole members of the Dapoxyl family produced by Molecular Probes are excellent agents for selective labeling of the endoplasmic reticulum in living cells, either alone or in combination with other dyes (90). These probes are retained after fixation with formaldehyde, but can be lost with permeabilizing detergents. Another useful probe is Brefeldin A (131), a stereochemically complex fungal metabolite that serves as an inhibitor of protein trafficking out of the endoplasmic reticulum. Finally, similar to other organelles, monoclonal antibodies (127–129) have been developed that target the endoplasmic reticulum in fixed cells for immunocytochemistry investigations.

QUANTUM DOTS

Nanometer-sized crystals of purified semiconductors known as quantum dots are emerging as a potentially useful fluorescent labeling agent for living and fixed cells in both traditional widefield and laser scanning confocal fluorescence microscopy (132–136). Recently introduced techniques enable the purified tiny semiconductor crystals to be coated with a hydrophilic polymer shell and conjugated to antibodies or other biologically active peptides and carbohydrates for application in many of the classical immunocytochemistry protocols (Fig. 15). These probes have significant benefits over organic dyes and fluorescent proteins, including long-term photostability, high fluorescence intensity levels, and multiple colors with single-wavelength excitation for all emission profiles (136).

Quantum dots produce illumination in a manner similar to the well-known semiconductor light emitting diodes, but are activated by absorption of a photon rather than an electrical stimulus. The absorbed photon creates an electron-hole pair that quickly recombines with the concurrent emission of a photon having lower energy. The most useful semiconductor discovered thus far for producing biological quantum dots is cadmium selenide (CdSe), a material in

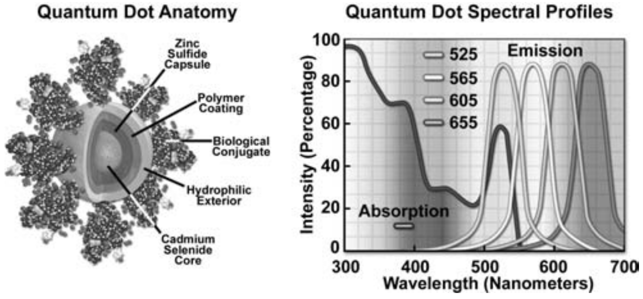


Figure 15. Anatomy and spectral profiles of quantum dot conjugates. The cadmium selenide core is encapsulated with zinc sulfide, and then a polymer coating is applied followed by a hydrophilic exterior to which the biological conjugate is attached (left). The absorption profile displays a shoulder at 400 nm, while the emission spectra all feature similar symmetrical profiles.

which the energy of the emitted photons is a function of the physical size of the nanocrystal particles. Thus, quantum dots having sizes that differ only by tenths of a nanometer emit different wavelengths of light, with the smaller sizes emitting shorter wavelengths, and vice versa.

Unlike typical organic fluorophores or fluorescent proteins, which display highly defined spectral profiles, quantum dots have an absorption spectrum that increases steadily with decreasing wavelength (Fig. 15). Also, in contrast, the fluorescence emission intensity is confined to a symmetrical peak with a maximum wavelength that is dependent on the dot size, but independent of the excitation wavelength (135). As a result, the same emission profile is observed regardless of whether the quantum dot is excited at 300, 400, 500, or 600 nm, but the fluorescence intensity increases dramatically at shorter excitation wavelengths. For example, the extinction coefficient for a typical quantum dot conjugate that emits in the orange region (605 nm) is approximately five-fold higher when the semiconductor is excited at 400 versus 600 nm. The fwhm value for a typical quantum dot conjugate is ~ 30 nm (135), and the spectral profile is not skewed towards the longer wavelengths (having higher intensity tails), such is the case with most organic fluorochromes. The narrow emission profile enables several quantum dot conjugates to be simultaneously observed with a minimal level of bleed through.

For biological applications, a relatively uniform population of cadmium selenide crystals is covered with a surrounding semiconductor shell composed of zinc sulfide to improve the optical properties. Next, the core material is coated with a polymeric film and other ligands to decrease hydrophobicity and to improve the attachment efficiency of conjugated macromolecules. The final product is a biologically active particle that ranges in size from 10 to 15 nm, somewhere in the vicinity of a large protein (133). Quantum dot conjugates are solubilized as a colloidal suspension in common biological buffers and may be incorporated into existing labeling protocols in place of classical staining reagents (such as organic fluorochrome-labeled secondary antibodies).

In confocal microscopy, quantum dots are excited with varying degrees of efficiency by most of the spectral lines produced by the common laser systems, including the argon-ion, helium–cadmium, krypton–argon, and the green helium–neon. Particularly effective at exciting quantum dots in the UV and violet regions are the new blue diode and diode-pumped solid-state lasers that have prominent spectral lines at 442 nm and below (135,136). The 405 nm blue diode laser is an economical excitation source that is very effective for use with quantum dots due to their high extinction coefficient at this wavelength. Another advantage of using these fluorophores in confocal microscopy is the ability to stimulate multiple quantum dot sizes (and spectral colors) in the same specimen with a single excitation wavelength, making these probes excellent candidates for multiple labeling experiments (137).

The exceptional photostability of quantum dot conjugates is of great advantage in confocal microscopy when optical sections are being collected. Unlike the case of organic fluorophores, labeled structures situated away from the focal plane do not suffer from excessive photobleaching during repeated raster scanning of the specimen and yield more accurate 3D volume models. In widefield fluorescence microscopy, quantum dot conjugates are available for use with conventional dye-optimized filter combinations that are standard equipment on many microscopes. Excitation can be further enhanced by substituting a shortpass filter for the bandpass filter that accompanies most filter sets, thus optimizing the amount of lamp energy that can be utilized to excite the quantum dots. Several of the custom fluorescence filter manufacturers offer combinations specifically designed to be used with quantum dot conjugates.

FLUORESCENT PROTEINS

Over the past few years, the discovery and development of naturally occurring fluorescent proteins and mutated derivatives have rapidly advanced to center stage in the investigation of a wide spectrum of intracellular processes in living organisms (75,78,80). These biological probes have provided scientists with the ability to visualize, monitor, and track individual molecules with high spatial and temporal resolution in both steady-state and kinetic experiments. A variety of marine organisms have been the source of >100 fluorescent proteins and their analogs, which arm the investigator with a balanced palette of noninvasive biological probes for single, dual, and multispectral fluorescence analysis (75). Among the advantages of fluorescent proteins over the traditional organic and new semiconductor probes described above is their response to a wider variety of biological events and signals. Coupled with the ability to specifically target fluorescent probes in subcellular compartments, the extremely low or absent photodynamic toxicity, and the widespread compatibility with tissues and intact organisms, these biological macromolecules offer an exciting new frontier in live-cell imaging.

The first member of this series to be discovered, GFP, was isolated from the North Atlantic jellyfish, *Aequorea Victoria*, and found to exhibit a high degree of fluorescence without the aid of additional substrates or coenzymes (138–142). In native green fluorescent protein, the fluorescent moiety is a tripeptide derivative of serine, tyrosine, and glycine that requires molecular oxygen for activation, but no additional cofactors or enzymes (143). Subsequent investigations revealed that the GFP gene could be expressed in other organisms, including mammals, to yield fully functional analogs that display no adverse biological effects (144). In fact, fluorescent proteins can be fused to virtually any protein in living cells using recombinant complementary DNA cloning technology, and the resulting fusion protein gene product expressed in cell lines adapted to standard tissue culture methodology. Lack of a need for cell-specific activation cofactors renders the fluorescent proteins much more useful as generalized probes than other biological macromolecules, such as the phycobiliproteins, which require insertion of accessory pigments in order to produce fluorescence.

Mutagenesis experiments with green fluorescent protein have produced a large number of variants with improved folding and expression characteristics, which have eliminated wild-type dimerization artifacts and fine tuned the absorption and fluorescence properties. One of the earliest variants, known as enhanced green fluorescence protein (EGFP), contains codon substitutions (commonly referred to as the S65T mutation) that alleviates the temperature sensitivity and increases the efficiency of GFP expression in mammalian cells (145). Proteins fused with EGFP can be observed at low light intensities for long time periods with minimal photobleaching. Enhanced green fluorescent protein fusion products are optimally excited by the 488 nm spectral line from argon and krypton–argon ion lasers in confocal microscopy. This provides an excellent biological probe and instrument combination for examining intracellular protein pathways along with the structural dynamics of organelles and the cytoskeleton.

Additional mutation studies have uncovered GFP variants that exhibit a variety of absorption and emission characteristics across the entire visible spectral region, which have enabled researchers to develop probe combinations for simultaneous observation of two or more distinct fluorescent proteins in a single organism (see the spectral profiles in Fig. 16). Early investigations yielded the blue fluorescent protein (BFP) and cyan fluorescent protein (CFP) mutants from simple amino acid substitutions that shifted the absorption and emission spectral profiles of wild-type GFP to lower wavelength regions (146–148). Used in combination with GFP, these derivatives are useful in resonance energy transfer (FRET) experiments and other investigations that rely on multicolor fluorescence imaging (73). Blue fluorescent protein can be efficiently excited with the 354 nm line from a high power argon laser, while the more useful cyan derivative is excited by a number of violet and blue laser lines, including the 405 nm blue diode, the 442 nm helium–cadmium spectral line, and the 457 nm line from the standard argon-ion laser.

Another popular fluorescent protein derivative, the yellow fluorescent protein (YFP), was designed on the basis of the GFP crystalline structural analysis to red-shift the absorption and emission spectra (148). Yellow fluorescent protein is optimally excited by the 514 nm spectral line of the argon-ion laser, and provides more intense emission than enhanced green fluorescent protein, but is more sensitive to low pH and high halogen ion concentrations. The enhanced yellow fluorescent protein derivative (EYFP) is useful with the 514 argon-ion laser line, but can also be excited with relatively high efficiency by the 488 nm line from argon and krypton–argon lasers. Both of these fluorescent protein derivatives have been widely applied to protein–protein FRET investigations in combination with CFP, and in addition, have proven useful in studies involving multiprotein trafficking.

Attempts to shift the absorption and emission spectra of *Aequorea Victoria* fluorescent proteins to wavelengths in the orange and red regions of the spectrum have met with little success. However, fluorescent proteins from other marine species have enabled investigators to extend the available spectral regions to well within the red wavelength range. The DsRed fluorescent protein and its derivatives, originally isolated from the sea anemone *Discosoma striata*, are currently the most popular analogs for fluorescence analysis in the 575–650 nm region (149). Another protein, HcRed from the *Heteractis crispa* purple anemone, is also a promising candidate for investigations in the longer wavelengths of the visible spectrum (150). Newly developed photoactivation fluorescent proteins, including photoactivatable green fluorescent protein (PA-GFP;74), Kaede (76), and kindling fluorescent protein 1 (KFP1; 151), exhibit dramatic improvements over GFP (up to several 1000-fold) in fluorescence intensity when stimulated by violet laser illumination. These probes should prove useful in fluorescence confocal studies involving selective irradiation of specific target regions and the subsequent kinetic analysis of diffusional mobility and compartmental residency time of fusion proteins.

QUENCHING AND PHOTOBLEACHING

The consequences of quenching and photobleaching are suffered in practically all forms of fluorescence microscopy, and result in an effective reduction in the levels of emission (152,153). These artifacts should be of primary

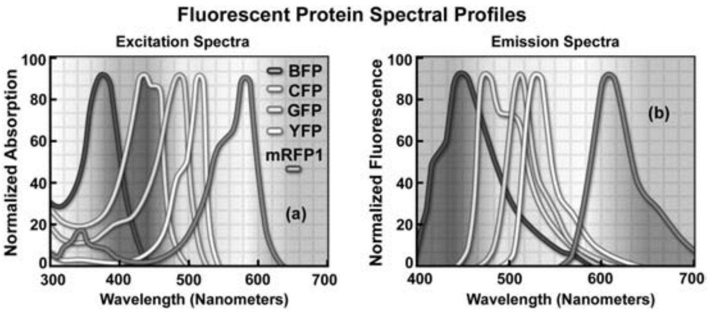


Figure 16. Fluorescent spectral profiles, plotted as normalized absorption or emission as a function of wavelength, for fluorescent proteins emitting in the blue to orange-red regions of the visible spectrum. Each profile is identified with a colored bullet in (a), which illustrates excitation spectra. (b) The emission spectra for the proteins according to the legend in (a).

consideration when designing and executing fluorescence investigations. The two phenomena are distinct in that quenching is often reversible whereas photobleaching is not (154). Quenching arises from a variety of competing processes that induce nonradiative relaxation (without photon emission) of excited-state electrons to the ground state, which may be either intramolecular or intermolecular in nature. Because nonradiative transition pathways compete with the fluorescence relaxation, they usually dramatically lower or, in some cases, completely eliminate emission. Most quenching processes act to reduce the excited state lifetime and the quantum yield of the affected fluorophore.

A common example of quenching is observed with the collision of an excited state fluorophore and another (nonfluorescent) molecule in solution, resulting in deactivation of the fluorophore and return to the ground state. In most cases, neither of the molecules is chemically altered in the collisional quenching process. A wide variety of simple elements and compounds behave as collisional quenching agents, including oxygen, halogens, amines, and many electron-deficient organic molecules (154). Collisional quenching can reveal the presence of localized quencher molecules or moieties, which via diffusion or conformational change, may collide with the fluorophore during the excited state lifetime. The mechanisms for collisional quenching include electron transfer, spin–orbit coupling, and intersystem crossing to the excited triplet state (154,155). Other terms that are often utilized interchangeably with collisional quenching are internal conversion and dynamic quenching.

A second type of quenching mechanism, termed static or complex quenching, arises from nonfluorescent complexes formed between the quencher and fluorophore that serve to limit absorption by reducing the population of active, excitable molecules (154,156). This effect occurs when the fluorescent species forms a reversible complex with the quencher molecule in the ground state, and does not rely on diffusion or molecular collisions. In static quenching, fluorescence emission is reduced without altering the excited state lifetime. A fluorophore in the excited state can also be quenched by a dipolar resonance energy transfer mechanism when in close proximity with an acceptor molecule to which the excited-state energy can be transferred nonradiatively. In some cases, quenching can occur through non molecular mechanisms, such as attenuation of incident light by an absorbing species (including the chromophore itself).

In contrast to quenching, photobleaching (also termed fading) occurs when a fluorophore permanently loses the ability to fluoresce due to photon-induced chemical damage and covalent modification (153–156). Upon transition from an excited singlet state to the excited triplet state, fluorophores may interact with another molecule to produce irreversible covalent modifications. The triplet state is relatively long lived with respect to the singlet state, thus allowing excited molecules a much longer timeframe to undergo chemical reactions with components in the environment (155). The average number of excitation and emission cycles that occur for a particular fluorophore before photobleaching is dependent on the molecular structure and the local environment (154,156). Some fluorophores bleach quickly after emitting only a few photons, while others that are more robust can undergo thousands or even millions of cycles before bleaching.

Figure 17 presents a typical example of photobleaching (fading) observed in a series of digital images captured at different time points for a multiply stained culture of normal Tahr ovary (HJ1.Ov line) fibroblast cells. The nuclei were stained with DAPI (blue fluorescence), while the mitochondria and actin cytoskeleton were stained with MitoTracker Red CMXRos (red fluorescence) and an Alexa Fluor phalloidin derivative (Alexa Fluor 488; green fluorescence), respectively. Time points were taken in 2 min intervals using a fluorescence filter combination with bandwidths tuned to excite the three fluorophores simultaneously while also recording the combined emission signals. Note that all three fluorophores have a relatively high intensity in Fig. 17a, but the DAPI (blue) intensity starts to drop rapidly at two min and is almost completely gone at six min (Fig. 17f). The mitochondrial and actin stains are more resistant to photobleaching, but the intensity of both drops dramatically over the course of the timed sequence (10 min).

An important class of photobleaching events is represented by events that are photodynamic, meaning they involve the interaction of the fluorophore with a combination of light and oxygen (157–161). Reactions between fluorophores and molecular oxygen permanently destroy fluorescence and yield a free-radical singlet oxygen species that can chemically modify other molecules in living cells. The amount of photobleaching due to photodynamic events is a function of the molecular oxygen concentration and the proximal distance between the fluorophore, oxygen molecules, and other cellular components. Photobleaching can be reduced by limiting the exposure time of

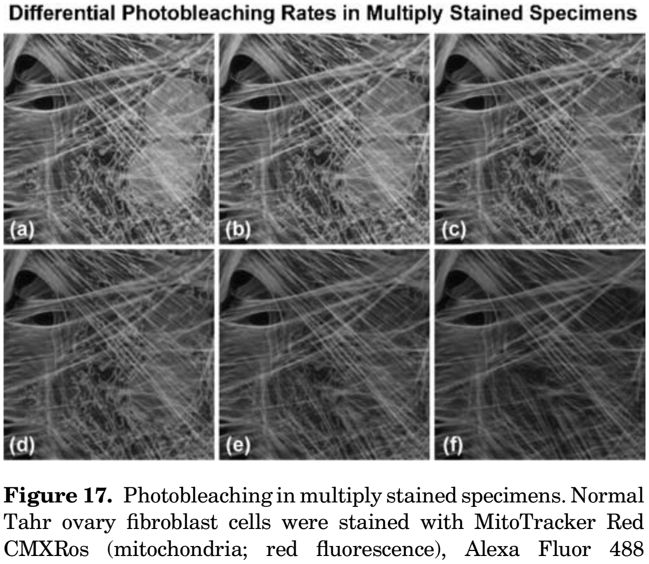

Differential Photobleaching Rates in Multiply Stained Specimens
(a)
(b)
(c)
(d)
(e)
(f)

Figure 17. Photobleaching in multiply stained specimens. Normal Tahr ovary fibroblast cells were stained with MitoTracker Red CMXRos (mitochondria; red fluorescence), Alexa Fluor 488 conjugated to phalloidin (actin; green fluorescence), and subsequently counterstained with DAPI (nuclei; blue fluorescence). Time points were taken in two-minute intervals over a 10 min period using a fluorescence filter combination with bandwidths tuned to excite the three fluorophores simultaneously while also recording the combined emission signals. (a–f) Time = 0, 2, 4, 6, 8, 10 min, respectively.

fluorophores to illumination or by lowering the excitation energy. However, these techniques also reduce the measurable fluorescence signal. In many cases, solutions of fluorophores or cell suspensions can be deoxygenated, but this is not feasible for living cells and tissues. Perhaps the best protection against photobleaching is to limit exposure of the fluorochrome to intense illumination (using neutral density filters) coupled with the judicious use of commercially available antifade reagents that can be added to the mounting solution or cell culture medium (153).

Under certain circumstances, the photobleaching effect can also be utilized to obtain specific information that would not otherwise be available. For example, in FRAP experiments, fluorophores within a target region are intentionally bleached with excessive levels of irradiation (82). As new fluorophore molecules diffuse into the bleached region of the specimen (recovery), the fluorescence emission intensity is monitored to determine the lateral diffusion rates of the target fluorophore. In this manner, the translational mobility of fluorescently labeled molecules can be ascertained within a very small (2–5 μm) region of a single cell or section of living tissue.

Although the subset of fluorophores that are advantageous in confocal microscopy is rapidly growing, many of the traditional probes that have been useful for years in widefield applications are still of little utility when constrained by fixed-wavelength laser spectral lines. Many of the limitations surrounding the use of fluorophores excited in the ultraviolet region will be eliminated with the introduction of advanced objectives designed to reduce aberration coupled to the gradual introduction of low cost, high power diode laser systems with spectral lines in these shorter wavelengths. The 405 nm blue diode laser is a rather cheap alternative to more expensive ion and Noble gas based ultraviolet lasers, and is rapidly becoming available for most confocal microscope systems. Helium–neon lasers with spectral lines in the yellow and orange region have rendered some fluorophores useful that were previously limited to widefield applications. In addition, new diode-pumped solid-state lasers are being introduced with emission wavelengths in the UV, violet, and blue regions.

Continued advances in fluorophore design, dual-laser scanning, multispectral imaging, endoscopic instruments, and spinning disk applications will also be important in the coming years. The persistent problem of emission crossover due to spectral overlap, which occurs with many synthetic probes and fluorescent proteins in multicolor investigations, benefits significantly from spectral analysis and deconvolution of lambda stacks. Combined, these advances and will dramatically improve the collection and analysis of data obtained from complex fluorescence experiments in live-cell imaging.

BIBLIOGRAPHY

1. Pawley JB, editor. Handbook of Biological Confocal Microscopy. New York: Plenum Press; 1995.
2. Paddock SW, editor. Confocal Microscopy: Methods and Protocols. Totowa (NJ): Humana Press; 1999.
3. Diaspro A, editor. Confocal and Two-Photon Microscopy: Foundations, Applications, and Advances. New York: Wiley-Liss; 2002.
4. Matsumoto B, editor. Cell Biological Applications of Confocal Microscopy. Methods in Cell Biology, Vol. 70. New York: Academic Press; 2002.
5. Sheppard CJR, Shotton DM. Confocal Laser Scanning Microscopy. Oxford (UK): BIOS Scientific Publishers; 1997.
6. Müller M. Introduction to Confocal Fluorescence Microscopy. Maastricht, The Netherlands: Shaker; 2002.
7. Hibbs AR. Confocal Microscopy for Biologists. New York: Kluwer Academic; 2004.
8. Conn PM. Confocal Microscopy. Methods in Enzymology, Vol. 307. New York: Academic Press; 1999.
9. Corle TR, Kino GS. Confocal Scanning Optical Microscopy and Related Imaging Systems. New York: Academic Press; 1996.
10. Wilson T, editor. Confocal Microscopy. New York: Academic Press; 1990.
11. Gu M. Principles of Three-Dimensional Imaging in Confocal Microscopes. New Jersey: World Scientific; 1996.
12. Masters BR, editor. Selected Papers on Confocal Microscopy. SPIE Milestone Series, Vol. MS 131. Bellingham (WA): SPIE Optical Engineering Press; 1996.
13. Mason WT. Fluorescent and Luminescent Probes for Biological Activity. New York: Academic Press; 1999.
14. Peterman EJG, Sosa H, Moerner WE. Single-Molecule Fluorescence Spectroscopy and Microscopy of Biomolecular Motors. Ann Rev Phys Chem 2004;55:79–96.
15. Goldman RD, Spector DL. Live Cell Imaging: A Laboratory Manual. New York: Cold Spring Harbor Press; 2005.
16. Minsky M. Microscopy Apparatus. US Patent 3,013,467. 1961.
17. Minsky M. Memoir on Inventing the Confocal Scanning Microscopy. Scanning 1988;10:128–138.
18. Egger MD, Petran M. New Reflected-Light Microscope for Viewing Unstained Brain and Ganglion Cells. Science 1967;157:305–307.
19. Davidovits P, Egger MD. Photomicrography of Corneal Endothelial Cells *in vivo*. Nature (London) 1973;244:366–367.
20. Amos WB, White JG. How the Confocal Laser Scanning Microscope entered Biological Research. Biol Cell 2003;95:335–342.
21. Brakenhoff GJ, Blom P, Barends P. Confocal Scanning Light Microscopy with High Aperture Immersion Lenses. J Micros 1979;117:219–232.
22. Sheppard CJR, Wilson T. Effect of Spherical Aberration on the Imaging Properties of Scanning Optical Microscopes. Appl Opt 1979;18:1058.
23. Hamilton DK, Wilson T. Scanning Optical Microscopy by Objective Lens Scanning. J Phys E: Sci Instr 1986;19:52–54.
24. Spring KR, Inoué S. Video Microscopy: The Fundamentals. New York: Plenum Press; 1997.
25. Wilson T. Optical Sectioning in Confocal Fluorescence Microscopes. J Micros 1989;154:143–156.
26. Lichtmann JW. Confocal Microscopy. Sci Am Aug. 1994; 40–45.
27. White JG, Amos WB, Fordham M. An Evaluation of Confocal versus Conventional Imaging of Biological Structures by Fluorescence Light Microscopy. J Cell Biol 1987;105:41–48.
28. Swedlow JR, et al. Measuring Tubulin Content in Toxoplasma gondii: A Comparison of Laser-Scanning Confocal and Wide-Field Fluorescence Microscopy. Proc Natl Acad Sci USA 2002;99:2014–2019.
29. Stelzer EHK. Practical Limits to Resolution in Fluorescence Light Microscopy. In: Yuste R, Lanni F, Konnerth A, editors. Imaging Neurons: A Laboratory Manual. New York: Cold Spring Harbor Press; 2000. pp 12.1–12.9.
30. Rost FWD. Fluorescence Microscopy. Vol. 1. New York: Cambridge University Press; 1992.

31. Murray J. Confocal Microscopy, Deconvolution, and Structured Illumination Methods. In: Goldman RD, Spector DL, editors. Live Cell Imaging: A Laboratory Manual. New York: Cold Spring Harbor Press; 2005. pp 239–280.
32. Dickinson ME, et al. Multi-Spectral Imaging and Linear Unmixing Add a Whole New Dimension to Laser Scanning Fluorescence Microscopy. Biotechniques 2001;31:1272–1278.
33. Zimmermann T, Rietdorf J, Pepperkok R. Spectral Imaging and its Applications in Live Cell Microscopy. FEBS Lett 2003;546:87–92.
34. Lansford R, Bearman G, Fraser SE. Resolution of Multiple Green Fluorescent Protein Variants and Dyes using Two-Photon Microscopy and Imaging Spectroscopy. J Biomed Opt 2001;6:311–318.
35. Hiraoka Y, Shimi T, Haraguchi T. Multispectral Imaging Fluorescence Microscopy for Living Cells. Cell Struct Funct 2002;27:367–374.
36. Gu Y, Di WL, Kelsell DP, Zicha D. Quantitative Fluorescence Energy Transfer (FRET) Measurement with Acceptor Photobleaching and Spectral Unmixing. J Microsc 2004;215:162–173.
37. Ford BK, et al. Computed Tomography-Based Spectral Imaging for Fluorescence Microscopy. Biophys J 2001;80:986–993.
38. Bach H, Renn A, Wild UP. Spectral Imaging of Single Molecules. Single Mol 2000;1:73–77.
39. Wilson T, Carlini AR. Three-Dimensional Imaging in Confocal Imaging Systems with Finite Sized Detectors. J Microsc 1988;149:51–66.
40. Murphy DB. Fundamentals of Light Microscopy and Electronic Imaging. New York: Wiley-Liss; 2001.
41. Wright SJ, Wright DJ. Introduction to Confocal Microscopy. In: Matsumoto B, editor. Cell Biological Applications of Confocal Microscopy. Methods in Cell Biology, Vol. 70. New York: Academic Press; 2002. pp 1–85.
42. Webb RH. Confocal Optical Microscopy. Rep Prog Phys 1996;59:427–471.
43. Wilhelm S, Gröbler B, Gluch M, Hartmut H. Confocal Laser Scanning Microscopy: Optical Image Formation and Electronic Signal Processing. Jena, Germany: Carl Zeiss Advanced Imaging Microscopy; 2003.
44. Ichihara A, et al. High-Speed Confocal Fluorescence Microscopy using a Nipkow Scanner with Microlenses for 3-D Imaging of Fluorescent Molecules in Real-Time. Bioimages 1996;4:57–62.
45. Inoué S, Inoué T. Direct-View High-Speed Confocal Scanner—The CSU-10. In: Matsumoto B, editor. Cell Biological Applications of Confocal Microscopy. Methods in Cell Biology, Vol. 70. New York: Academic Press; 2002. p 88–128.
46. Nakano A. Spinning-Disk Confocal Microscopy —A Cutting-Edge Tool for Imaging of Membrane Traffic. Cell Struct Funct 2002;27:349–355.
47. Chong FK, et al. Optimization of Spinning Disk Confocal Microscopy: Synchronization with the Ultra-Sensitive EMCCD. In: Conchello JA, Cogswell CJ, Wilson T, editors. Three-Dimensional and Multidimensional Microscopy: Image Acquisition and Processing XI. Proc. SPIE. 2004; 5324: 65–76.
48. Sandison D, Webb W. Background Rejection and Signal-to-Noise Optimization in the Confocal and Alternative Fluorescence Microscopes. Appl Op 1994;33:603–610.
49. Centonze VE, White JG. Multiphoton Excitation Provides Optical Sections from Deeper within Scattering Specimens than Confocal Imaging. Biophys J 1998;75:2015.
50. Boccacci P, Bertero M. Image-Restoration Methods: Basics and Algorithms. In: Diaspro A, editor. Confocal and Two-Photon Microscopy: Foundations, Applications, and Advances. New York: Wiley-Liss; 2002. pp 253–269.
51. Verveer PJ, Gemkow MJ, Jovin TM. A Comparison of Image Restoration Approaches Applied to Three-Dimensional Confocal and Wide-Field Fluorescence Microscopy. J Micros 1998;193:50–61.
52. Conchello JA, Hansen EW. Enhanced 3-D Reconstruction from Confocal Scanning Microscope Images. 1: Deterministic and Maximum Likelihood Reconstructions. Appl Op 1990;29: 3795–3804.
53. Al-Kofahi O, et al. Algorithms for Accurate 3D Registration of Neuronal Images Acquired by Confocal Scanning Laser Microscopy. J Micros 2003;211:8–18.
54. Conchello JA, et al., editors. Three-Dimensional and Multidimensional Microscopy: Image Acquisition and Processing. Vol. I–XII, Bellingham (WA): SPIE International Society for Optical Engineering; 1994–2005.
55. Centonze V, Pawley J. Tutorial on Practical Confocal Microscopy and use of the Confocal Test Specimen. In: Pawley JB, editor. Handbook of Biological Confocal Microscopy. New York: Plenum Press; 1995. p 549–570.
56. Gratton E, vandeVen MJ. Laser Sources for Confocal Microscopy. In: Pawley JB, editor. Handbook of Biological Confocal Microscopy. New York: Plenum Press; 1995. p 69–98.
57. Ashkin A, Dziedzic JM, Yamane T. Optical Trapping and Manipulation of Single Cells using Infrared Laser Beams. London 1987;330:769–771.
58. DeMaggio S. Running and Setting Up a Confocal Microscope Core Facility. In: Matsumoto B, editor. Cell Biological Applications of Confocal Microscopy. Methods in Cell Biology, Vol. 70. New York: Academic Press; 2002. p 475–486.
59. Spring KR. Detectors for Fluorescence Microscopy. In: Periasamy A, editor. Methods in Cellular Imaging. New York: Oxford University Press; 2001. p 40–52.
60. Art J. Photon Detectors for Confocal Microscopy. In: Pawley JB, editor. Handbook of Biological Confocal Microscopy. New York: Plenum Press; 1995. p 183–196.
61. Amos WB. Instruments for Fluorescence Imaging. In: Allan VJ, editor. Protein Localization by Fluorescence Microscopy: A Practical Approach. New York: Oxford University Press; 2000. pp 67–108.
62. Hergert E. Detectors: Guideposts on the Road to Selection. Photonics Design and Applications Handbook. 2001. pp H110–H113.
63. Piston DW, Patterson GH, Knobel SM. Quantitative Imaging of the Green Fluorescent Protein (GFP). In: Sullivan KF, Kay SA, editors. Green Fluorescent Proteins, Methods in Cell Biology, Vol. 58. New York: Academic Press; 1999. pp 31–47.
64. Carter DR. Photomultiplier Handbook: Theory, Design, Application. Lancaster (PA): Burle Industries, Inc.; 1980.
65. Cody SH, et al. A Simple Method Allowing DIC Imaging in Conjunction with Confocal Microscopy. J Microsc 2005;217: 265–274.
66. Chang IC. Acousto-Optic Devices and Applications. In: Bass M, Van Stryland EW, Williams DR, Wolfe WL, editors. Optics II: Fundamentals, Techniques, and Design. New York: McGraw-Hill; 1995. pp 12.1–12.54.
67. Wachman ES. Acousto-Optic Tunable Filters for Microscopy. In: Yuste R, Lanni F, Konnerth A, editors. Imaging Neurons: A Laboratory Manual. New York: Cold Spring Harbor Press; 2000. p 4.1–4.8.
68. Shonat RD, et al. Near-Simultaneous Hemoglobin Saturation and Oxygen Tension Maps in Mouse Brain using an AOTF Microscope. Biophy J 1997;73:1223–1231.
69. Wachman ES, Niu W, Farkas DL. Imaging Acousto-Optic Tunable Filter with 0.35-Micrometer Spatial Resolution. App Opt 1996;35:5220–5226.
70. Wachman ES. AOTF Microscope for Imaging with Increased Speed and Spectral Versatility. Biophys J 1997;73:1215–1222.

71. Chen Y, Mills JD, Periasamy A. Protein Localization in Living Cells and tissues using FRET and FLIM. Differentiation 2003;71:528–541.

72. Wallrabe H, Periasamy A. Imaging Protein Molecules using FRET and FLIM Microscopy. Curr Opin Biotech 2005;16: 19–27.

73. Day RN, Periasamy A, Schaufele F. Fluorescence Resonance Energy Transfer Microscopy of Localized Protein Interactions in the Living Cell Nucleus. Methods 2001;25:4–18.

74. Patterson GH, Lippincott-Schwartz J. A Photoactivatable GFP for Selective Photolabeling of Proteins and Cells. Science 2002;297:1873–1877.

75. Verkhusha VV, Lukyanov KA. The Molecular Properties and Applications of Anthozoa Fluorescent Proteins and Chromoproteins. Nature Biotechnol 2004;22:289–296.

76. Ando R, et al. An Optical Marker Based on the UV-Induced Green-to-Red Photoconversion of a Fluorescent Protein. Proc Natl Acad Sci USA 2002;99:12651–12656.

77. Sharma D. The Use of an AOTF to Achieve High Quality Simultaneous Multiple Label Imaging. Bio-Rad Technical Notes. San Francisco: Bio-Rad, Note 4; 2001.

78. Miyawaki A, Sawano A, Kogure T. Lighting up Cells: Labeling Proteins with Fluorophores. Nature Cell Biol 2003;5:S1–S7.

79. Zhang J, Campbell RE, Ting AY, Tsien RY. Creating New Fluorescent Probes for Cell Biology. Nature Rev Mol Cell Bio 2002;3:906–918.

80. Lippincott-Schwartz J, Patterson G. Development and Use of Fluorescent Protein Markers in Living Cells. Science 2003;300:87–91.

81. Klonis N, et al. Fluorescence Photobleaching Analysis for the Study of Cellular Dynamics. Eur Biophys J 2002;31:36–51.

82. Lippincott-Schwartz J, Altan-Bonnet N, Patterson GH. Photobleaching and Photoactivation: Following Protein Dynamics in Living Cells. Nature Cell Biol 2003;5:S7–S14.

83. Phair RD, Misteli T. Kinetic Modelling Approaches to *in vivo* Imaging. Nature Rev Mol Cell Bio 2002;2:898–907.

84. Politz JC. Use of Caged Fluorophores to Track Macromolecular Movement in Living Cells. Trends Cell Biol 1999;9:284–287.

85. Born M, Wolf E. Principles of Optics. New York: Cambridge University Press; 1999.

86. Stelzer EHK. Contrast, Resolution, Pixelation, Dynamic range, and Signal-to-Noise Ratio: Fundamental Limits to Resolution in Fluorescence Light Microscopy. J Microsc 1997;189:15–24.

87. Pawley J. Fundamental Limits in Confocal Microscopy. In: Pawley JB, editor. Handbook of Biological Confocal Microscopy. New York: Plenum Press; 1995. p 19–37.

88. Webb RH, Dorey CK. The Pixelated Image. In: Pawley JB, editor. Handbook of Biological Confocal Microscopy. New York: Plenum Press; 1995. p 55–67.

89. Jonkman JEN, Stelzer EHK. Resolution and Contrast in Confocal and Two-Photon Microscopy. In: Diaspro A, editor. Confocal and Two-Photon Microscopy: Foundations, Applications, and Advances. New York: Wiley-Liss; 2002. p 101–125.

90. Haugland RP. The Handbook: A Guide to Fluorescent Probes and Labeling Technologies. Chicago: Invitrogen Molecular Probes; 2005.

91. Lemasters JJ, et al. Confocal Imaging of Ca^{2+}, pH, Electrical Potential and Membrane Permeability in Single Living Cells. Methods Enzymol 1999;302:341–358.

92. Johnson I. Fluorescent Probes for Living Cells. Histochem J 1998;30:123–140.

93. Kasten FH. Introduction to Fluorescent Probes: Properties, History, and Applications. In: Mason WT, editor. Fluorescent and Luminescent Probes for Biological Activity. New York: Academic Press; 1999. p 17–39.

94. Coons AH, Creech HJ, Jones RN, Berliner E. Demonstration of Pneumococcal Antigen in Tissues by use of Fluorescent Antibody. J Immunol 1942;45:159–170.

95. Tsien RY. Building and Breeding Molecules to Spy on Cells and Tumors. FEBS Lett 2005;579:927–932.

96. Bruchez Jr M, et al. Semiconductor Nanocrystals as fluorescent Biological Labels. Science 1998;218:2013–2016.

97. Tsien RY, Waggoner A. Fluorophores for Confocal Microscopy. In: Pawley JB, editor. Handbook of Biological Confocal Microscopy. New York: Plenum Press; 1995. p 267–280.

98. Wessendorf MW, Brelje TC. Which Fluorophore is Brightest? A Comparison of the Staining Obtained Using Fluorescein, Tetramethylrhodamine, Lissamine Rhodamine, Texas Red and Cyanine 3.18. Histochemistry 1992;98:81–85.

99. Entwistle A, Noble M. The use of Lucifer Yellow, BODIPY, FITC, TRITC, RITC and Texas Red for Dual Immunofluorescence Visualized with a Confocal Scanning Laser Microscope. J Microsc 1992;168:219–238.

100. Darzynkiewicz Z. Differential Staining of DNA and RNA in Intact Cells and Isolated Cell Nuclei with Acridine Orange. Methods Cell Biol 1990;33:285–298.

101. Waring MJ. Complex Formation Between Ethidium Bromide and Nucleic Acids. J Mol Biol 1965;13:269–282.

102. Arndt-Jovin DJ, Jovin TM. Fluorescence Labeling and Microscopy of DNA. Fluores Microsc Living Cells Culture Part B Methods Cell Biol 1989;30:417–448.

103. Kubista M, Akerman B, Norden B. Characterization of Interaction between DNA and 4′,6-Diamidino-2-phenylindole by Optical Spectroscopy. Biochemistry 1987;26:4545–4553.

104. Loewe H, Urbanietz J. Basic Substituted 2,6-Bisbenzimidazole Derivatives: A Novel Series of Substances with Chemotherapeutic Activity. Arzneim-Forsch 1974;24:1927–1933.

105. Arndt-Jovin DJ, Jovin TM. Analysis and Sorting of Living Cells According to Deoxyribonucleic Acid Content. J Histochem Cytochem 1977;25:585–589.

106. Durand RE, Olive PL. Cytotoxicity, Mutagenicity and DNA Damage by Hoechst 33342. J Histochem Cytochem 1982; 30:111–116.

107. Panchuk-Voloshina N, et al. Alexa Dyes, A Series of New Fluorescent Dyes that Yield Exceptionally bright, Photostable Conjugates. J Histochem Cytochem 1999;47:1179–1188.

108. Berlier JE, et al. Quantitative Comparison of Long-Wavelength Alexa Fluor Dyes to Cy Dyes: Fluorescence of the Dyes and their Conjugates. J Histochem Cytochem 2003;51:1699–1712.

109. Mujumdar RB, et al. Cyanine Dye Labeling Reagents: Sulfoindocyanine Succinimidyl Esters. Bioconjugate Chem 1993;4:105–111.

110. Ballou B, et al. Tumor Labeling *in vivo* using Cyanine-Conjugated Monoclonal Antibodies. Cancer Immunol Immunother 1995;41:257–263.

111. Zorov DB, Kobrinsky E, Juhaszova M, Sollott SJ. Examining Intracellular Organelle Function Using Fluorescent Probes. Circul Res 2004;95:239–252.

112. Stephens DG, Pepperkok R. The Many Ways to Cross the Plasma Membrane. Proc Natl Acad Sci USA 2001;98:4295–4298.

113. Rudolf R, Mongillo M, Rizzuto R, Pozzan T. Looking Forward to Seeing Calcium. Nature Rev Mol Cell Biol 2003;4:579–586.

114. Martin H, Bell MG, Ellis-Davies GC, Barsotti RJ. Activation Kinetics of Skinned Cardiac Muscle by Laser Photolysis of Nitrophenyl-EGTA. Biophys J 2004;86:978–990.

115. White C, McGeown G. Imaging of Changes in Sarcoplasmic Reticulum [Ca^{2+}] using Oregon Green BAPTA 5N and Confocal Laser Scanning Microscopy. Cell Calcium 2002;31:151–159.

116. Helm PJ, Patwardhan A, Manders EM. A Study of the Precision of Confocal, Ratiometric, Fura-2-Based [Ca^{2+}] Measurements. Cell Calcium 1997;22:287–298.

117. Rijkers GT, Justement LB, Griffioen AW, Cambier JC. Improved Method for Measuring Intracellular Ca^{++} with Fluo-3. Cytometry 1990;11:923–927.

118. Schild D, Jung A, Schultens HA. Localization of Calcium Entry through Calcium Channels in Olfactory Receptor Neurons using a Laser Scanning Microscope and the Calcium Indicator Dyes Fluo-3 and Fura-Red. Cell Calcium 1994;15: 341–348.
119. Willoughby D, Thomas RC, Schwiening CJ. Comparison of Simultaneous pH Measurements made with 8-Hydroxypyrene-1,3,6-trisulphonic acid (HPTS) and pH-Sensitive Microelectrodes in Snail Neurons. Pflugers Arch 1998;436:615–622.
120. Ozkan P, Mutharasan R. A Rapid Method for Measuring Intracellular pH using BCECF-AM. Biochim Biophys Acta 2002;1572:143.
121. Cody SH, et al. Intracellular pH Mapping with SNARF-1 and Confocal Microscopy. I: A Quantitative Technique for Living Tissue and Isolated Cells. Micron 1993;24:573–580.
122. Dubbin PN, Cody SH, Williams DA. Intracellular pH Mapping with SNARF-1 and Confocal Microscopy. II: pH Gradients within Single Cultured Cells. Micron 1993;24:581–586.
123. Poot M, et al. Analysis of Mitochondrial Morphology and Function with Novel Fixable Fluorescent Stains. J Histochem Cytochem 1996;44:1363–1372.
124. Keij JF, Bell-Prince C, Steinkamp JA. Staining of Mitochondrial Membranes with 10-Nonyl Acridine Orange, MitoFluor Green, and MitoTracker Green is Affected by Mitochondrial Membrane Potential Altering Drugs. Cytometry 2000;39: 203–210.
125. Reers M. et al. Mitochondrial Membrane Potential Monitored by JC-1 Dye. Methods Enzymol 1995;260:406–417.
126. Price OT, Lau C, Zucker RM. Quantitative Fluorescence of 5-FU-Treated Fetal Rat Limbs using Confocal Laser Scanning Microscopy and LysoTracker Red. Cytometry 2003;53A:9–21.
127. Kumar RK, Chapple CC, Hunter N. Improved Double Immunofluorescence for Confocal Laser Scanning Microscopy. J Histochem Cytochem 1999;47:1213–1217.
128. Suzuki T, Fujikura K, Higashiyama T, Takata K. DNA Staining for Fluorescence and Laser Confocal Microscopy. J Histochem Cytochem 1997;45:49–53.
129. Haugland RP. Coupling of Monoclonal Antibodies with Fluorophores. In: Davis WC, editor. Monoclonal Anibody Protocols, Methods in Molecular Biology. Vol. 45. Totowa, (NJ): Humana Press; 1995. p 205–221.
130. Pagano RE, Martin OC. Use of Fluorescent Analogs of Ceramide to Study the Golgi Apparatus of Animal Cells. In: Celis JE, editor. Cell Biology: A Laboratory Handbook. Vol. 2, 1998. p 507–512.
131. Cole L, Davies D, Hyde GJ, Ashford AE. ER-Tracker Dye and BODIPY-Brefeldin A Differentiate the Endoplasmic Reticulum and Golgi Bodies from the Tubular-Vacuole System in Living Hyphae of Pisolithus tinctorius. J Microsc 2000;197:239–249.
132. Jaiswal JK, Mattoussi H, Mauro JM, Simon SM. Long-Term Multiple Color Imaging of Live Cells using Quantum Dot Bioconjugates. Nature Biotechnol 2003;21:47–52.
133. Larson DR, et al. Water Soluble Quantum Dots for Multiphoton Fluorescence Imaging *in vivo*. Science 2003;300: 1434–1436.
134. Watson A, Wu X, Bruchez M. Lighting up Cells with Quantum Dots. Biotechniques 2003;34:296–303.
135. Michalet X, et al. Quantum Dots for Live Cells, in vivo Imaging, and Diagnostics. Science 2005;307:538–544.
136. Gao X, et al. *In vivo* Molecular and Cellular Imaging with Quantum Dots. Curr Opin Biotech 2005;16:63–72.
137. Lacoste TD, et al. Ultrahigh-Resolution Multicolor Colocalization of Single Fluorescent Probes. Proc Natl Acad Sci USA 2000;97:9461–9466.
138. Tsien RY. The Green Fluorescent Protein. Ann Rev Biochem 1998;67:509–544.
139. Sullivan KF, Kay SA, editors. Green Fluorescent Proteins, Methods in Cell Biology. Vol. 58. New York: Academic Press; 1999.
140. Conn PM, editor. Green Fluorescent Protein, Methods in Enzymology. Vol. 302. New York: Academic Press; 1999.
141. Hicks BW, editor. Green Fluorescent Protein, Methods in Molecular Biology. Vol. 183. Totowa (NJ): Humana Press; 2002.
142. Chalfie M, Kain S, editors. Green Fluorescent Protein: Properties, Applications, and Protocols. New York: Wiley-Liss; 1998.
143. Zimmer M. Green Fluorescent Protein: Applications, Structure, and Related Photophysical Behavior. Chem Rev 2002;102:759–781.
144. Chalfie M, et al. Green Fluorescent Protein as a Marker for Gene Expression. Science 1994;263:802–805.
145. Heim R, Cubitt AB, Tsien RY. Improved Green Fluorescence. Nature (London) 1995;373:664–665.
146. Heim R, Prasher DC, Tsien RY. Wavelength Mutations and Posttranslational Autoxidation of Green Fluorescent Protein. Proc Natl Acad Sci USA 1994;91:12501–12504.
147. Heim R, Tsien RY. Engineering Green Fluorescent Protein for Improved Brightness, Longer Wavelengths, and Fluorescence Resonance Energy Transfer. Curr Biol 1996;6: 178–182.
148. Wachter RM, et al. Structural Basis of Spectral Shifts in the Yellow-Emission Variants of Green Fluorescent Protein. Structure 1998;6:1267–1277.
149. Matz MV, et al. Fluorescent Proteins from Nonbioluminescent Anthozoa Species. Nature Biotechnol 1999;17:969–973.
150. Matz MV, Lukyanov KA, Lukyanov SA. Family of the Green Fluorescent Protein: Journey to the End of the Rainbow. BioEssays 2002;24:953–959.
151. Natural Animal Coloration can be Determined by a Nonfluorescent Green Fluorescent Protein Homolog. J Biol Chem 2000;275:25879–25882.
152. Song L, Hennink EJ, Young IT, Tanke HJ. Photobleaching Kinetics of Fluorescein in Quantitative Fluorescence Microscopy. Biophys J 1995;68:2588–2600.
153. Berrios M, Conlon KA, Colflesh DE. Antifading Agents for Confocal Fluorescence Microscopy. Methods Enzymol 1999; 307:55–79.
154. Lakowicz JR. Principles of Fluorescence Spectroscopy. New York: Kluwer Academic/Plenum Publishers; 1999.
155. Song L, Varma CA, Verhoeven JW, Tanke HJ. Influence of the Triplet Excited State on the Photobleaching Kinetics of Fluorescein in Microscopy. Biophys J 1996;70:2959–2968.
156. Herman B. Fluorescence Microscopy. New York: BIOS Scientific Publishers; 1998.
157. Bunting JR. A Test of the Singlet Oxygen Mechanism of Cationic Dye Photosensitization of Mitochondrial Damage. Photochem Photobiol 1992;55:81–87.
158. Byers GW, Gross S, Henrichs PM. Direct and Sensitized Photooxidation of Cyanine Dyes. Photochem Photobiol 1976; 23:37–43.
159. Dittrich PS, Schwille P. Photobleaching and Stabilization of Fluorophores used for Single Molecule Analysis with One- and Two-Photon Excitation. Appl Phys B 2001;73:829–837.
160. Gandin E, Lion Y, Van de Vorst A. Quantum Yield of Singlet Oxygen Production by Xanthene Derivatives. Photochem Photobiol 1983;37:271–278.
161. Kanofsky JR, Sima PD. Structural and Environmental Requirements for Quenching of Singlet Oxygen by Cyanine Dyes. Photochem Photobiol 2000;71:361–368.

Further Reading

Willison JR. Signal Detection and Analysis, In: Bass, M. Van Stryland, E. W. Williams DR, Wolfe WL. Optics I: Fundamentals, Techniques, and Design. New York: McGraw-Hill; pp 1995. 18.1–18.16.

See also CELLULAR IMAGING; FLUORESCENCE MEASUREMENTS; ION-SENSITIVE FIELD EFFECT TRANSISTORS.

MICROSCOPY, ELECTRON

Mahrokh Dadsetan
Lichun Lu
Michael J. Yaszemski
Mayo Clinic,
College of Medicine
Rochester, Minnesota

INTRODUCTION

Invention of the light microscope by Janssens in 1590 was the first milestone in the microscopic world. Janssens' microscope magnified objects up to 20–30 times their original size. By the beginning of the twentieth century, objects could be magnified only up to 1000 times with a resolution of 0.2 μm. In the early 1930s, the limitations of light microscopes and the scientific desire to see intracellular structural details, such as mitochondria and nuclei led to the development of electron microscopes. The electron microscope took advantage of the much shorter wavelength of the electron compared to that of visible light. With the electron microscope, another 1000-fold increase in magnification was accomplished with a concomitant increase in resolution, allowing visualization of viruses, deoxyribonuclic acid (DNA), and smaller objects, such as molecules and atoms. The transmission electron microscope (TEM) was the first type of electron microscope, and was developed by Ruska and Knoll in Germany in 1931. Electron microscopy is based on a fundamental physics concept stated in the de Broglie theory (1924). This concept is that moving electrons have the properties of waves. The second major advancement in electron microscopy was made by Busch, who demonstrated in 1926 that electrostatic or magnetic fields could be used as a lens to focus an electron beam. In 1939, Siemens Corp. began commercial production of a microscope developed by Von Borries and Ruska in Germany. Hiller, Vance and others constructed the first TEM in North America in 1941. This instrument had a resolution of 2.5 nm.

About the same time that the first TEM was nearing completion in the 1930s, a prototype of the scanning electron microscope (SEM) was constructed by Knoll and Von Ardenne in Germany. However, the resolution of this microscope was no better than that of the light microscope. Following several improvements made by RCA in the United States, as well as Cambridge University in England, a commercial SEM became available in 1963. A later version of the SEM made by the Cambridge Instrument Co. had a resolving power of ~20–50 nm and a useful magnification of 75,000×. Recent models of the SEM have a resolving power of 3.0 nm and magnifications up to 300,000×.

Although the design of TEM and SEM is similar in many ways, their applications are very different. The TEM is patterned after as light microscope, except that electrons instead of light pass through the object. The electrons are then focused by two or more electron lenses to form a greatly magnified image onto photographic film or a charge coupled device (CCD) camera. The image produced by TEM is two-dimensional (2D) and the brightness of a particular region of the image is proportional to the number of electrons that are transmitted through the specimen at that position on the image. The SEM produces a three-dimensional (3D) image by scanning the surface of a specimen with a 2–3 nm spot of electrons to generate secondary electrons from the specimen that are then detected by a sensor. The resolution of an SEM is limited by two quite different sets of circumstances. One of these is concerned with the physics of electron optics, while the other depends on the penetration of electrons into the object being imaged.

A third type of electron microscope, the scanning transmission electron microscope (STEM) has features of both the transmission and scanning electron microscopes. This microscope is an analytical tool that determines the presence and distribution of the atomic elements in the specimen. Recently, two groups of researchers have accomplished a subangstrom resolution (0.06 nm) for STEM using an aberration corrector. They have reported that columns of atoms in a silicone crystal that are 0.078 nm apart can be distinguished at this resolution (1). The image of Si shown in Fig. 1 has been recorded in a high angle annular dark field (HAADF) mode, and the pairs of atomic columns are seen directly resolved. The HAADF detector collects electrons scattered by the sample to angles greater than the detector inner radius. Such high angle scattering is largely incoherent thermal diffuse scattering, which means that the resolution observed in the image is determined by the intensity distribution of the illuminating probe. With this advantage over conventional coherent high resolution transmission electron microscopy (HRTEM), HAADF–STEM has enabled imaging not only of individual atomic columns in crystals, but single dopant atoms on their surface and within their interior.

In 1982, another type of electron microscope, the scanning tunneling microscope (STM) was developed by two scientists, Rohrer and Binnig, for studying surface structure. This invention was quickly followed by the development of a family of related techniques classified as scanning probe microscopy (SPM). These techniques are based upon moving a probe (typically called a tip in STM, which is literally a sharp metallic object) just above a specimen's surface while monitoring some interaction

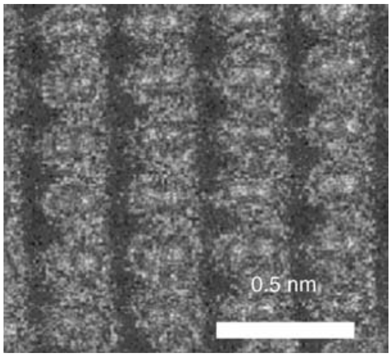


Figure 1. Image of a silicone crystal observed in the [112] orientation recorded with an aberration corrected STEM. (From Ref. 1). Reproduced by courtesy of American Association for the Advancement of Science.)

between the probe and the surface. Atomic force microscopy (AFM) is another important technique of this kind but is not categorized as electron microscopy. Bennig and Rohrer were awarded one-half of the 1986 Nobel Prize in physics for invention of the STM, while Ernst Ruska was awarded the other one-half of that same Nobel Prize for his 1931 invention of the electron microscope. In STM, electrons are speeded up in a vacuum until their wavelength is extremely short, only one hundred-thousandth that of white light. Beams of these fast-moving electrons are focused on a cell sample and are absorbed or scattered by the cell's parts so as to form an image on an electron-sensitive photographic plate. The STM is widely used in both industrial and academic research to obtain atomic scale images of metal surfaces. It provides a 3D profile of the surface, which is useful in characterizing surface roughness and determining the size and conformation of surface molecules. (2) Invention of electron microscopy had an enormous impact in the field of biology, specifically in cell and tissue analysis. Almost 15 years after the invention of the first electron microscope by Ruska, many efforts were made to apply this technique to biological problems. Using the electron microscope, cell organelles and cell inclusions were discovered or resolved in finer details. Electron microscopy, specifically TEM, is now among the most important tools in cell biology and diagnostic pathology.

The latest advancement in electron microscopy is 3D reconstruction of cellular components at a resolution that is on the order of magnitude of atomic structures defined by X-ray crystallography. The method for reconstruction of 3D images of single, transparent objects recorded by TEM is called electron tomography (ET). In order to generate 3D images of individual molecules, one needs to obtain as many tilt images as possible, covering the widest possible angular range. The representative images of particles obtained from different orientations is then analyzed and combined by a software program to reconstruct the molecule in 3D. With improvements in instrumentation, data collection methods and techniques for computation, ET may become a preferred method for imaging isolated organelles and small cells. So far, the electron tomography method covers the resolution range of 2.5–5.0 nm. Data obtained via electron tomography furnish a rich source of quantitative information about the structural composition and organization of cellular components. It offers the opportunity to obtain 3D information on structural cellular arrangements with a significantly higher resolution than that provided by any other method currently available (e.g., confocal laser microscopy) (3).

THEORY OF ELECTRON MICROSCOPY

According to electromagnetic theory, a light source initiates a vibrational motion that transmits energy in the direction of propagation. The wave motion of light is analogous to that produced by a stone thrown into a pool of water. When the waves generated from throwing a stone strike an object that has an opening or aperture, another series of waves is generated from the edge of the object. The result is a new source of waves that emerges with the original waves. This bending or spreading phenomenon is known as diffraction. Diffracted waves interfere with the initial waves, and the result is an image of the edge of the object. The edge appears to have a series of bands or fringes called Fresnel fringes running parallel to the edge (Fig. 2). Thus, if a strong beam of light illuminates a pinhole in a screen and thus a pinhole serves as a point source and the light passing through is focused by an apertured "perfect" lens on a second screen, the image obtained is not a pinpoint of light, but rather a bright central disk surrounded by a diffuse ring of light. Even if monochromatic light was used to illuminate the point source and was to pass through a perfect lens, the image will not be a sharp one, but rather a diffuse disk composed of concentric rings. This type of image is known as an Airy disk after Sir George Airy, who first described this pattern during the nineteenth century (Fig. 3) (4). To determine resolving power (RP), it is important to know the radius of the Airy disk. The radius of the Airy disk as measured to the first dark ring (r) is expressed by following equation:

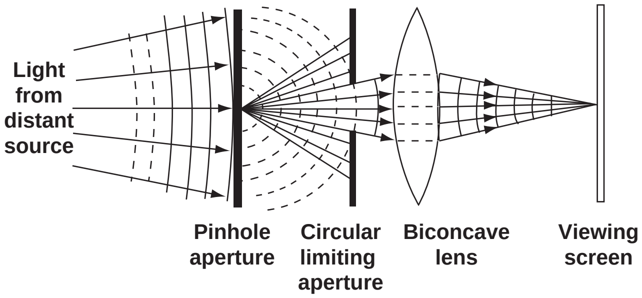


Figure 2. Diffraction of light waves.

$$r = \frac{0.612\,\lambda}{n(\sin\alpha)} \tag{1}$$

In Eq. 1, λ = wavelength of illumination; n = refractive index of the medium between the point source and the lens, relative to free space; α = half the angle of the cone of light from the specimen plane accepted by the front lens of objective

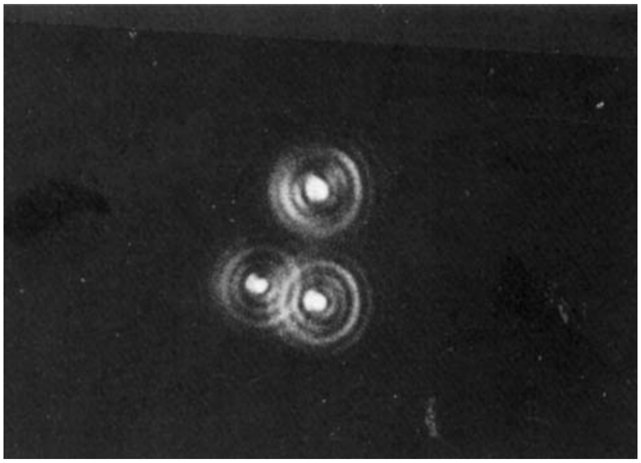


Figure 3. Airy disks generated by viewing three pinholes in a light microscope. Magnification of micrograph is 1000×. (From Ref. 4. Reproduced by courtesy of Jones and Bartlett Publishers.)

The above equation can be shown in another form as:

$$d = \frac{0.612\,\lambda}{\mathrm{NA}} \tag{2}$$

where NA (numerical aperture) = n sin α and represents the light gathering power of the lens aperture.

From the above equation, RP is defined as the minimum distance that two objects can be placed apart and still be seen as separate entities. Consequently, the shorter the distance, the better (or higher) is the RP of the system. For example, consider a light microscope using ultraviolet (UV) light, which lies beyond the lower end of the visible spectrum (400 nm). Further specifications of this system include a glass slide with standard immersion oil (refractive index, $n = 1.5$), and sinα= 0.87 (sine of a 64° angle, representing one-half of the 128° acceptance angle of a glass lens. The theoretical resolution that can be attained by this system is ~ 0.2 μm. In other words, two points in the specimen that are not separated by at least this distance will not be seen as two distinct points, but will be observed as a single blurred image. Since the values of sin α and n cannot be significantly increased beyond the stated values, the RP can most effectively be improved by reducing wavelength.

Electron Beams And Resolution

The concept that moving electrons might be used as an illumination source was suggested by a tenet of the de Broglie theory, that moving electrons have wave properties. The wavelength of this particle-associated radiation is given by following equation:

$$\lambda \, \frac{h}{mv} \tag{3}$$

where m is the mass of the particle, v the velocity of particle, and h is Planck's constant ($6.626 \times 10^{-34}\mathrm{J}^{-1}$). For an electron accelerated by a potential of 60,000 V (60 kV), the wavelength of the electron beam would be ~0.005 nm, which is 100,000 times shorter than that for green light. By using Eq. 1, a TEM with perfect lenses would therefore in theory be able to provide a resolution of 0.0025 nm. In practice, the actual resolution of a modern high resolution transmission electron microscope is closer to 0.2 nm. The reason we are not able to achieve the nearly 100-fold better resolution of 0.002 nm is due to extremely narrow aperture angles (~1000 times smaller than that of the light microscope) needed by the electron microscope lenses to overcome a major resolution limiting phenomenon called spherical aberration. In addition, diffraction, chromatic aberration and astigmatism all contribute to decreased resolution in TEM, and need to be corrected to achieve higher resolution. (5)

Magnification

The maximum magnification of any microscope is simply the ratio of the microscope's resolution to the resolution of the unaided human eye. The resolution of the eye viewing an object at 25 cm is generally taken to be 0.25 mm. Since the resolution of a light microscope is ~0.25 μm, maximum useful light magnification is ~1000×, obtainable from an objective lens of 100× followed by an eyepiece of 10×. The magnification of TEM would be ~0.25mm/0.25 nm. This is a 10^6× magnification, and corresponds to a 1000-fold increase in resolution compared to a light microscope. An objective lens of 100× is followed by an "intermediate" lens of 25×, and the final image is projected by a projector lens of 100×. Further magnification for critical focusing is obtained by viewing the image on the fluorescent screen with a long working distance binocular microscope of 10×. The final image is photographed at 250,000×. The processed negative is then enlarged a further 4× in a photographic enlarger. This result in a final prints (the electron micrograph) at the desired magnification of 10^6×6.

Electromagnetic Lenses

An electromagnetic lens is generated by a coil of wire with a direct current (dc) that passes through the coil. This electromagnetic coil is called a solenoid. It forms an axially and radially symmetric magnetic field that converges to a point. A divergent cone of electrons enters from a point source, and thus forms a real image on the lens axis. An advantage of electromagnetic lenses is that the focal length can be made infinitely variable by varying the coil current. Therefore, both magnification and image focus can be adjusted by controlling the lens current (Fig. 4) (7).

Lens Aberrations

Electron lenses are affected by all the aberrations of optical lenses, such as spherical aberration, chromatic aberration, astigmatism, and distortion. Spherical aberration results from the geometry of both glass and electromagnetic lenses such that rays passing through the periphery of the lens are refracted more than rays passing along the axis. Spherical aberration may be reduced by using an aperture to eliminate some of the peripheral rays. Although this aperture is attractive for reducing spherical aberration, it decreases the aperture angle and thereby prevents the electron microscope from achieving the theoretical resolution predicted by Eq. 1.

Chromatic aberration results when electromagnetic radiations of different energies converge at different focal

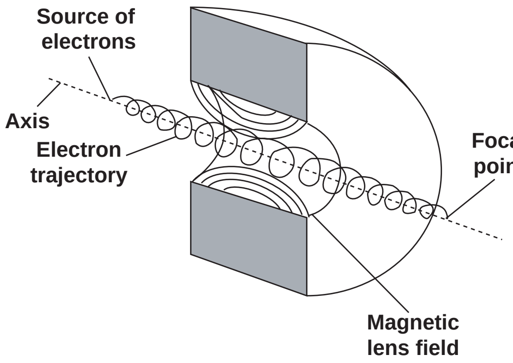


Figure 4. Single electron passing through electromagnetic lens. The electron is focused by the magnetic field to follow a trajectory that will converge at a defined focal point after it emerges from the lens.

planes. Chromatic aberration results in the enlargement of a focal point with a consequential loss of resolution. It can be corrected by using a monochromatic source of electromagnetic radiation. This entails stabilizing the accelerating voltage to generate the electrons with same levels of energy, and having a good vacuum to minimize the energy loss of the electrons during their passage through the transmission specimen. This effect can also be reduced by decreasing the aperture of the objective lens. (6)

Astigmatism is caused by radial asymmetry in a lens, giving rise to a focal length in one plane that is different from that in another plane. The fabrication and maintenance of a lens that has a perfectly symmetric lens field is not feasible in practice. Thus, it is necessary to correct astigmatism by applying a radial symmetry compensator device called a stigmator. This consists of an adjustable electric or magnetic field that can be applied across the lens in any chosen direction, thus compensating for astigmatism. Image distortion, due to magnification changing across the field from the value at the center, may be positive (called barrel distortion) or negative (called pincushion distortion). These effects can be compensated by operating two lenses in series, arranging for barrel distortion in one to be compensated by pincushion distortion in the other. The lens system in modern electron microscopes is designed to automatically counterbalance the various types of distortions throughout a wide magnification range. (4)

DESIGN OF THE TRANSMISSION ELECTRECTRON MICROSCOPE

Both the light and electron microscopes are similar so far as the arrangement and function of their components are concerned. Thus, both microscopes, when used for photographic purposes, can be conveniently divided into the following component systems.

Illuminating System

This system serves to produce the required radiation and to direct it onto the specimen. It consists of source and condenser lenses.

Source Lens. The source of electrons, or cathode, is a hairpin of fine tungsten wire about 2 mm long, maintained at ~2500 K by ~2 W of alternting current (ac) or dc power. Electrons boil off the white-hot tungsten surface, and are shaped into a conical beam by an electrode system called the gun. Two further electrodes, the shield and the anode, combine to form an electrostatic collimating lens and accelerator. A suitable accelerating voltage (20–100 kV) is chosen for the specimen under examination, and is applied to the cathode as a negative potential so that the anode may remain at earth potential. The cathode and shield are therefore carried on an insulator. The filament-shield voltage (cathode bias) is made variable to adjust the total current drawn from the filament, which in turn varies the brightness of the final image (4).

The energy of the electrons in the TEM determines the relative degree of penetration of electrons into a specific sample, or alternatively, influences the thickness of material from which useful information may be obtained. Thus, a high energy TEM (400 kV) not only provides the highest resolution but also allows for the observation of relatively thick samples (e.g., ~0.2 μm) when compared with the more conventional 100 kV or 200 kV instruments. Because of the high spatial resolution obtained, TEMs are often employed to determine the detailed crystallography of fine-grained, or rare, materials (6,8).

Condenser Lens. The condenser lens regulates the convergence (and thus the intensity) of the illuminating beam on the specimen. The divergent electron beam emerging from the anode aperture can, in simple instruments, be used to illuminate the specimen directly. However, sufficient image brightness for high magnification is difficult to obtain. As in the light microscope, a condenser system is almost invariably interposed between the gun and specimen to concentrate the beam on the region of the specimen under examination. A single condenser lens suffices for electronoptical work up to 50,000×. However, for high resolution work a double condenser system is always used, which will concentrate the beam into an area as small as 1μ diameter.

Specimen Manipulation System

The pierced metal grid carrying the specimen proper is clamped at its periphery to a suitable holder designed to conduct heat rapidly away. The specimen temperature in a TEM may rise to 200 °C. The holder, including its attached specimen grid is introduced into the evacuated specimen chamber through an airlock by means of an insertion tool. This tool is then generally withdrawn after the holder has been placed on the translation stage. The holder is then free to move with the stage, which is driven from outside the column through airlocks, by means of levers and micrometer screws. Two mutually perpendicular stage movements, each of about ±1 mm, allow any part of the grid area to be brought to the microscope axis and viewed at the highest magnification. Most instruments provide a scan magnification so that the whole grid area may be viewed at ~100×. Suitable specimen areas are then chosen, and centered for study at higher magnifications.

Imaging System

This part of the microscope includes the objective, intermediate, and projector lenses. It is involved in the generation of the image and the magnification and projection of the final image onto a viewing screen or camera system. Electrons transmitted by the specimen enter the objective lens. Those passing through the physical aperture are imaged at ~100× in the intermediate lens object plane, 10–20 cm below the specimen. The position of this primary image plane is controlled by the objective lens current (focus control). A second image, which may be magnified or diminished, is formed by the intermediate lens, the current through which controls overall magnification (magnification control). This secondary image, formed in the objective plane of the projector lens, is then further magnified, and the overall magnification is determined by the position of the fluorescent screen or film.

Image Recording System

The final image is projected onto a viewing screen coated with a phosphorescent zinc-activated cadmium sulfide powder. This powder is attached to the screen with a binder, such as cellulose nitrate. Most electron microscopes provide for an inclination of the viewing screen so that the image may be conveniently examined either with the unaided eye or through a stereomicroscope (the binoculars). Although the stereomicroscope image may appear to be rough due to the 100 μm sized phosphorescent particles that make up the screen, it is necessary to view a magnified image in order to focus accurately. Some microscopes may provide a second, smaller screen that is brought into position for focusing. In this case, the main screen remains horizontal, except during exposure of the film. All viewing screens will have areas marked to indicate where to position the image so that it will be properly situated on the film. Preevacuated films are placed into an air lock (*camera chamber*) under the viewing screen and the chamber evacuated to high vacuum. The chamber is then opened to the column to permit exposure of the film. In modern electron microscopes (Fig. 5), exposure is controlled by an electrically operated shutter placed below the projector lens. As one begins to raise the viewing screen, the shutter blocks the beam until the screen is in the appropriate position for exposure. The shutter is then opened for the proper interval, after which the beam is again blocked until the screen is repositioned.

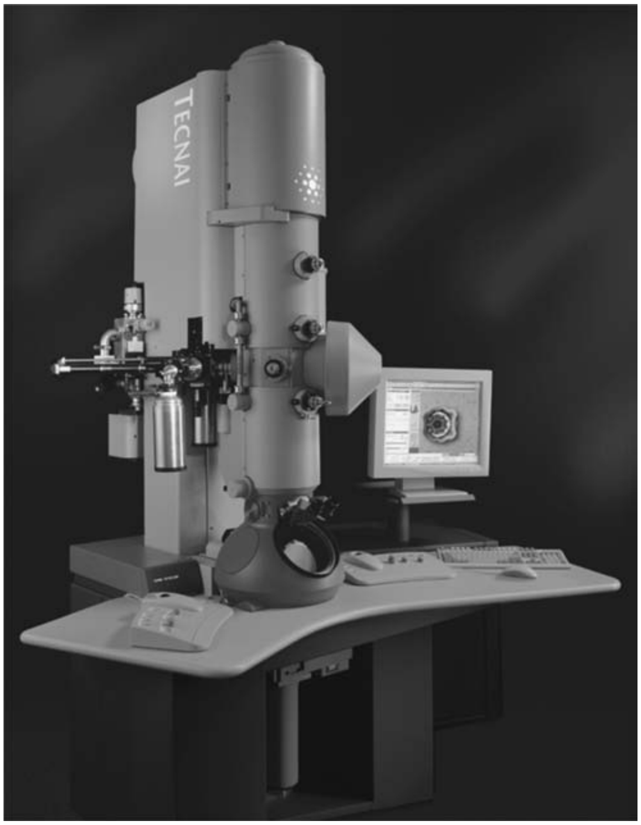


Figure 5. Image of a 300 kV TEM (FEI-Tecnai G^2 Polara) for cryoapplications at liquid nitrogen and liquid helium temperatures. (Reproduced by courtesy of FEI Company.)

DESIGN OF THE SCANNING ELECTRON MICROSCOPE

The SEM is made up of two basic systems, and the specimen is at their boundary. The first system is the electron optical column that provides the beam of illumination that is directed to the specimen. The second system consists of the electron collection, signal amplification, and image display units, which converts the electrons emitted from the specimen into a visible image of the specimen.

Electron Optical Column

The electron gun and electron lenses are present in the electron optical column of the SEM in an analogous fashion to their presence in the TEM.

1. Electron Gun: The electron source is most commonly the hairpin tungsten filament located in a triode electron gun. The electrons are emitted by the filament (also called the cathode), and accelerated by a field produced by the anode. The anode is usually at a positive potential on the order of 15 kV with respect to the cathode. A third electrode, the shield, lies between the anode and cathode and is negative with respect to the cathode. After leaving the bias shield and forming an initial focused spot of electrons of ~50 μm in diameter, a series of two to three condenser lenses are used to successively demagnify this spot sometimes down to ~2 nm. These small spot sizes are essential for the resolutions required at high magnifications. A heated tungsten filament is the conventional electron source for most SEMs; other special sources are lanthanum hexaboride (LaB_6) and the field emission guns (FEG). Both of these latter sources produce bright beams of small diameter and have much longer lifetimes than heated tungsten filaments. Schottky emission has largely replaced earlier source technologies based on either tungsten and LaB6 emission or cold-field emission in today's focused electron beam equipment including SEM, TEM, Auger systems, and semiconductor inspection tools. Schottky and cold-field emission are superior to thermionic sources in terms of source size, brightness and lifetime. Both are up to 1000 times smaller and up to 100 times brighter than thermionic emitters.
2. Electron Lenses: Most SEMs have three magnetic lenses in their column: the first, second, and final condenser lenses. The first condenser lens begins the demagnification of the 50 μm focused spot of electrons formed in the region of the electron gun. As the amount of current running through the first condenser lens is increased, the focal length of the lens becomes progressively shorter and the focused spot of electrons becomes smaller. In our earlier discussion of electron lenses, it was noted that focusing takes place by varying the focal length. This is accomplished by changing the intensity of the lens coil current, which in turn alters the intensity of the magnetic field that is generated by the lens. As the lens current increases, the lens strength increases

and the focal length decreases. A short focal length lens consequently causes such a wide divergence of the electrons leaving the lens that many electrons are not able to enter the next condenser lens. The overall effect of increasing the strength of first condenser lens is to decrease the spot size, but with a loss of electrons. An aperture is positioned in the lenses to decrease the spot size and reduce spherical aberration by excluding the more peripheral electrons. Each of the condenser lenses behaves in a similar manner and possesses apertures.

In designing the final condenser lens, several performance characteristics must be considered:

(a) Aberrations. Since the intermediate images of the crossover produced by the condenser lenses have significantly larger diameters than the final spot size, the effect of aberrations on these lenses are relatively small. It is thus the effects of spherical and chromatic aberration as well as the astigmatism of the final condenser lens that are critical in the design and performance of the objective lens of SEM.

(b) Magnetic Field. As a result of electron bombardment, secondary electrons in the SEM are emitted over a wide solid angle. These have energies of only few electron volts, yet they must be able to reach the detector to produce the necessary signal. As a result, the magnetic field at the specimen must be designed so that it will not restrict effective secondary electron collection.

(c) Focal Length. The extent of lens aberrations is dependent upon the focal length. Thus, it is desirable to keep the latter as short as possible in order to help minimize the effects of aberrations.

The final lens usually has externally adjustable apertures. Normally, final apertures on the order of 50–70 μm are used to generate smaller, less electron dense spots for secondary electron generation and imaging. Larger apertures, for example, 200 μm, are used to generate larger spots with greater numbers of electrons. These large spots contain a great deal of energy and may damage fragile specimens. They are used primarily to generate X rays for elemental analysis rather than for imaging purposes.

Specimen Manipulation System

The specimen is normally secured to a metal stub and is grounded to prevent the build up of static high voltage charges when the beam electrons strike the specimen. In order to orient the specimen precisely, relative to the electron beam and electron detectors, all SEMs have controls for rotating and traversing the specimen in x, y, and z directions. It is also possible to tilt the specimen in order to enhance the collection of electrons by a particular detector. These movements have a large effect on magnification, contrast, resolution and depth of field. Some improvement can be made in imaging by reorientation of the specimen.

Interaction Of Electron Beam With Specimen

Three basic possibilities exist as to the nature of the beam–specimen interaction used to generate the image:

1. Some primary electrons, depending on the accelerating voltage, penetrate the solid to depths as much as 10 μm. The electrons scatter randomly throughout the specimen until their energy is dissipated by interaction with atoms of the specimen.
2. Some primary electrons collide with or pass close to the nucleus of an atom of the specimen such that there is a change in the electron's momentum. This results in electron scatter through a large angle and electron reflection from the specimen. Such elastically reflected primary electrons are known as backscattered electrons.
3. Some primary electrons interact with the host atoms so that as a result of collisions, a cascade of secondary electrons is formed along the penetration path. Secondary electrons have energy ranges of 0–50 eV and are the electrons most commonly used to generate the 3D image. The mean path length of secondary electrons in many materials is $\sim$1 nm. Thus, although electrons are generated throughout the region excited by the incident beam, only those electrons that originate $<$1 nm deep in the sample escape to be detected as secondary. The shallow depth of production of detected secondary electrons makes them very sensitive to topography.

In addition to producing backscattered and secondary electrons, specimen–beam interactions also produce photons, specimen currents, Auger electrons and X rays that are characteristic of the probed specimen. These emanations can be detected by X ray or electron spectroscopy for elemental analysis of the specimen surface. However, it is rarely used for biological specimens.

Signal Versus Noise

The signals generated as a result of the electron beam striking a specimen are used to convey different types of information about the specimen. In the usual SEM imaging mode, signals consist of the secondary electrons generated from the spot struck by the electron beam and noise consists of secondary electrons originating at locations away from where the beam struck the specimen. The image quality is eventually expressed by the signal-to-noise (S/N) ratio. In a poor quality image the signal to noise ratio is low. One may achieve a better image by either reducing the noise or raising the signal. Since it is more difficult to reduce the noise level, the signal is usually raised by increasing the electron emissions from the gun. Several methods to accomplish increased electron emissions include: altering the bias settings, decreasing the distance between the anode and the filament, decreasing the distance between the filament and the shield aperture, and using either a lanthanum hexaboride filament or a cold-field emissions gun. A second method to increase the signal is to use slower scan rates on the specimen. Longer dwell times of the beam on the specimen will generate more secondary electrons from the spot where the beam strikes the specimen. This increase in current, however, carries with it an increased risk of damage to sensitive specimens (9).

Secondary Electron Detection

To collect the secondary electrons, a suitable electrode is held at a positive potential and serves to attract them and produce an emission current. The strength of this signal is proportional to the number of electrons striking the collector. This signal is used, after amplification; to modulate the intensity of the cathode-ray tube (CRT) beam as it moves across the tube face, synchronously with the path of the electron probe across the specimen surface.

Typically, the secondary electron collector is based on the original 1960 scintillator-photomultiplier design of Everhart and Thoronley. In this system, the secondary electrons are accelerated towards the scintillator by a potential difference of a few hundred to a few thousands volts. Upon hitting the scintillator, each electron produces many photons that are guided by the light pipe to the photomultiplier. Each photoelectron triggers a release of two or more secondary electrons at the first electrode (dynode) and process cascades, yielding from 100,000 to 50 million additional electrons. Thus, the photomultiplier reconverts the light to an electron current and provides a high degree of amplification that can be controlled by variation of the voltage applied to the dynodes (8).

Image Recording System

The final magnified image in the SEM is formed on a CRT or monitor. Unlike TEM, in which the electrons interact directly with the photographic medium, SEM images are most often photographed directly from the monitor through the lens of either a 35 mm roll film camera or a larger 4 in. × 5 in. (10.6 cm × 12.7 cm) sheet film camera. The camera shutter remains open as the electron beam slowly scans across the specimen. A valuable addition to most SEMs is the automatic data display that permits the generation of informational data on the viewing and recording monitors. With this accessory, experiment numbers, dates, accelerating voltages and magnifications may be displayed.

DIAGOSTIC ELECTRON MICROSCOPY

Electron microscopy excels as a diagnostic tool with respect to the detection and identification of both abnormal tissue anatomy and the pathogens responsible for the disease. The value of electron microscopy in difficult diagnostic situations has been demonstrated repeatedly, particularly when there is close coordination between the pathologist and the attending clinician. Ultrastructural study may be applied to a variety of substances including biological materials. By examination of specially prepared tissue sections, changes not perceived by light microscopy can be identified, leading to improved diagnostic interpretations. For example, in certain kidney diseases, such as nephrotic syndrome and Nil disease, the correct diagnosis can be made only by these means, and this in turn affects the selection of therapy. Similarly, certain neoplasms can be identified definitively only through ultrstructural studies, with obvious implications for treatment and prognosis. Another area of growing importance is the identification of viral particles in biological material. In some instances, ultrastructural study is the only way to establish the presence of a viral infection, and in other instances a diagnosis may be made earlier than by serological methods. The costs of diagnostic electron microscopy are relatively small in light of the benefits to patient care. The following are examples highlighting the use of electron microscopy in the diagnosis of certain diseases.

Neoplasms

Many neoplasms appear undifferentiated by light microscopy, but most show differentiation along one cell line or another at the ultrastructural level. However, it is noteworthy that not all of the ultrastructureal criteria for identifying the cell type may be present in every neoplasm. As expected, the more differentiated the neoplasm, the more likely will its cells contain a broad complement of diagnostic morphologic features. Usually, the ultrastructural findings do allow the pathologist to make a definitive diagnosis when interpreted in conjunction with the light microscopic picture, and in some cases with the histochemical and immunohistochemical results. The example that follows will highlight the use of diagnostic electron microscopy in the identification of carcinomas. Various types of carcinomas have a number of distinguishing features, but one common characteristic of all carcinomas is the presence of intracellular junctions, usually desmosomes and/or intermediate junctions. The presence of lumens, microvilli, tight junctions, junctional complexes, basal lamina, secretory granules, prominent Golgi apparatus, and moderately prominent rough endoplasmic reticulum are all suggestive of adenocarcinoma in the differential diagnosis of a neoplasm (10). Figure 6 shows a pancreatic carcinoma, which may arise from acinar cells, centrocinar cells, intercalated duct cells, interlobular duct cells, interlobular duct cells and main pancreatic duct cells. Adenocarcinomas arising from main and interlobular ducts (mucinous cystadenocarcinomas) have cells similar to those of bile ducts and intestinal epithelium; that is, the cytoplasm contains mucin granules, and the free surface has microvilli filled with thin filaments that anchor into the subjacent cytoplasm (11–13).

Infectious Diseases

Bacteria. Diagnostic criteria for bacterial rods and cocci are (*1*) the presence of an outer-cell wall, (*2*) the presence flagella or pili (fimbria) on the outer surface of the cell, (*3*) the presence of an inner-cell membrane, (*4*) a central nuclear region (nucleoid), without a limiting membrane, (*5*) dense cytoplasm composed mostly of ribosomes, and (*6*) a varying number of vesicles formed from the inner-cell membrane (mesosomes), storage vacuoles and endospores (14). Figure 7 shows the bacteria in Whipple disease. The rods are present both free and within macrophages.

Viruses. The distinct morphology of members of different viral families usually allows an agent to be assigned to

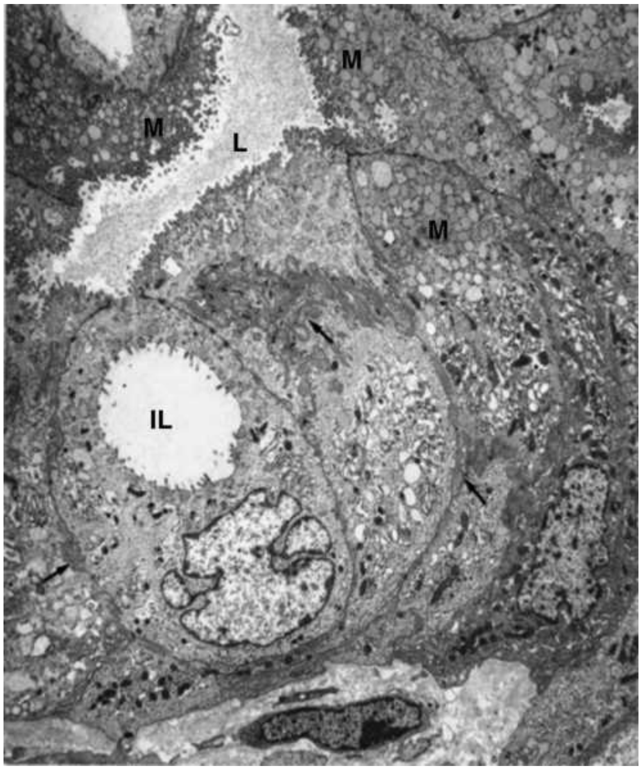


Figure 6. Ductal, mucinous cystadenocarcinoma (pancreas). In this field, the neoplastic cells form a cystic lumen (L) lined by innumerable microvilli. An intracytoplasmic lumen (IL), without junctional com- plexes, is present in one cell. Some of the cells lining the lumen have a rich collection of mucinous granules (M) in their apical cytoplasm. Lateral cell borders show a switch-backing pattern of interdigitation (arrows). (6800×) (From Ref. 10. Reproduced by courtesy of Springer-Verlag GmbH.)

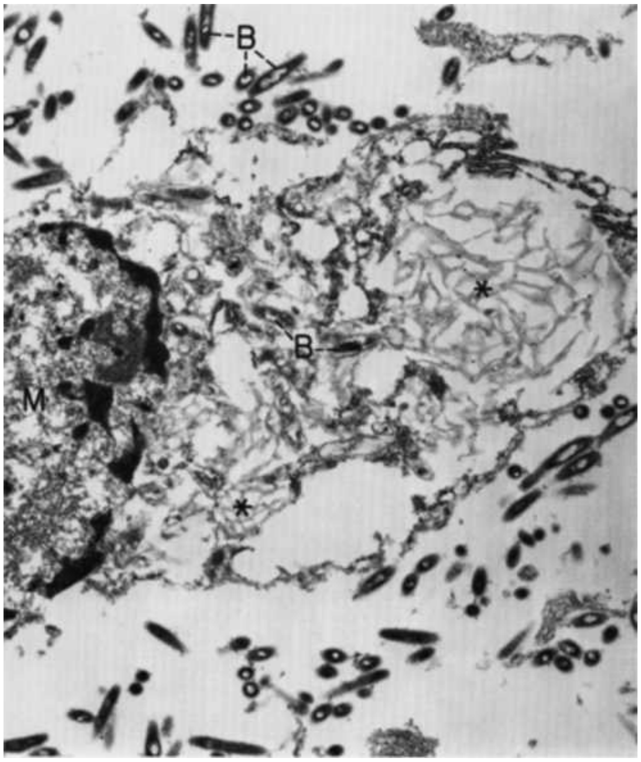


Figure 7. Whippl's disease. The lamina propria of the jejunal mucosa contained numerous Whipple's type macrophages [M= macrophage nucleus) and bacteria rods (B)]. Most of the intact bacilli are extracellular, whereas those in the macrophage are in various stages of degeneration, including the end-stage of serpiginous membrane (*). (16,500 ×) (From Ref. 10. Reproduced by courtesy of Springer-Verlag GmbH.)

a particular family. This morpho-diagnosis, combined with clinical information is often sufficient to permit a provisional diagnosis and to initiate treatment and containment protocols while waiting for other test results. Diagnostic criteria that apply to viruses in general are (*1*) the presence of intracellular and/or extracellular elliptical, stand-like, round or polygonal structures measuring 20–300 nm in diameter; and (*2*) the identification of viral morphology, consisting of a central, electron dense core (DNA-containing nucleoid) and an outer shell (capsid), which may have more than one layer (Fig. 8) (15).

Fungi. The electron microscopic diagnostic criteria for fungi include the identification of mononucleated oval yeast forms measuring 2–4 μm in diameter, with a thin cell wall and no true capsule. These can be located either extracellularly or intracellularly (16). A representative fungus, *Histoplasma capsulatum*, is shown in Fig. 9. The organisms have a clear halo between their visible cytoplasm and their thin cell wall.

Skeletal Muscle Diseases

The skeletal muscle responses to injury that are visible with the electron microscope can be categorized as follows: (*1*) alterations in the sarcolemma (e.g., discontinuities of the plasma membrane or the basement membrane); (*2*) alterations in myofilaments; (*3*) Z-band alterations (e.g., streaming and nemaline bodies); (*4*) nuclear changes (e.g., abnormal location of the nucleus within the muscle fiber and nuclear inclusions); (*5*) abnormalities of the sarcoplasmic reticulum and the T-system (e.g., tubular aggregates), (*6*) abnormal accumulations of metabolites (e.g., glycogen and lipids); (*7*) abnormal cytoplasmic structures (e.g., vacuoles, cytoplasmic bodies, concentric laminated bodies, fingerprint bodies, curvilinear bodies). In general, many of these ultrastructural abnormalities are not specific for a single disease. Electron microscopy can be a valuable adjunct to help the pathologist arrive at the proper interpretation of a muscle biopsy when taken together with all other available clinical, electrophysiologic, and histopathological data. In addition to the pathologic changes that might involve the muscle fibers themselves, many diseases of muscle also simultaneously affect adjoining connective tissue components, blood vessels and intramuscular nerves. It is therefore important to pay particular attention to these structures when examining muscle with the light and electron microscope (10). The light micrograph shown in Fig. 10 demonstrates centrally placed nuclei in the majority of the muscle fibers. The central nuclei often are surrounded by a clear area that is devoid of adenosine triphosphatase (ATPase) activity. Ultrastructural features of the paranuclear clear zone in Fig. 11 include (*1*) the

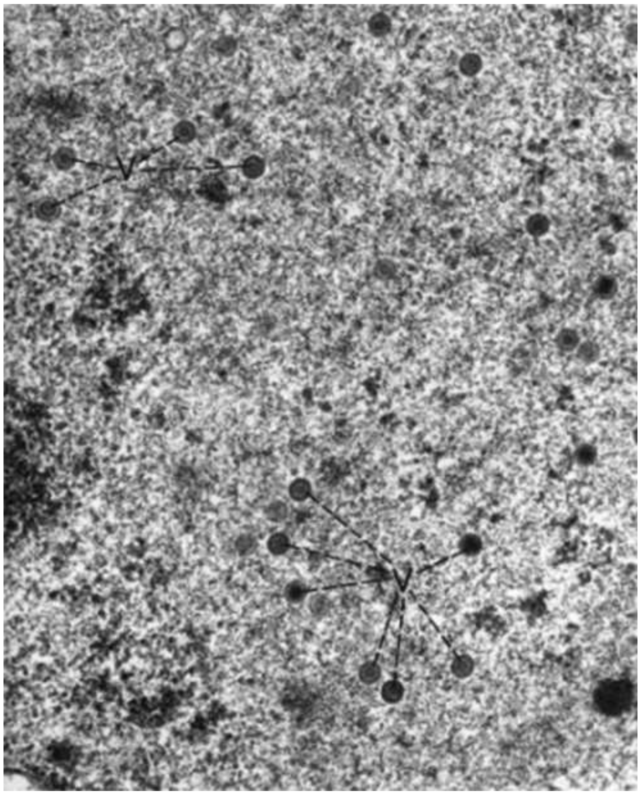


Figure 8. Herpes simplex encephalitis (cerebrum). Virions (V) have a central dense nucleoid and an outer three-layered capsid. (63,800×) (From Ref. 10. Reproduced by courtesy of Springer-Verlag GmbH.)

absence of myofilaments, (*2*) numerous mitochondria, (*3*) glycogen accumulation. In some patients, especially infants, the central clear zone may be more evident than the nuclei within that zone when examining the tissue in cross-section (10).

Peripheral Nerve Diseases

Wallerian Degeneration. Ultrastructural changes detected in peripheral nerve specimens include either the general pathologic responses of the peripheral nerve to either the injury or the specific disease entity that afflicts the patient. The general pathologic processes involving peripheral nerve can be divided into two broad categories: those that indicate a process primarily affecting the axon and those that indicate a process primarily affecting the myelin sheath. Examination of peripheral nerve biopsies by electron microscopy therefore must include evaluation of the axons, the interstitium, and the myelin and Schwann cells.

The sequence of structural changes following nerve injury that are collectively called Wallerian degeneration is shown in Fig. 12. Wallerian degeneration specifically refers to degeneration of the distal segments of a peripheral nerve after severance of the axons from their cell bodies (17). When the nerve injury is a contusion, the basement membrane of the Schwann cell is preserved, allowing regeneration within the endoneurial tube. In contrast, when the nerve injury is a transection, the endoneurial tube (composed of denervated Schawnn cells and extracellular matrix) may not be appropriately aligned with the regenerating axons. Axonal regeneration is therefore less efficient after nerve degeneration that follows a transection injury compared to that following a crush injury (10,17).

PROSPECTS OF ELECTRON MICROSCOPY

In the 1980s, electron microscopy lost much of its former role in the life sciences due to the introduction of modern, highly effective molecular analytical techniques, such as immunohistochemistry, chip technology, and confocal laser scan microscopy. In recent years, however, substantial technical improvements were made in specimen preparation, instrumentation and software, allowing the

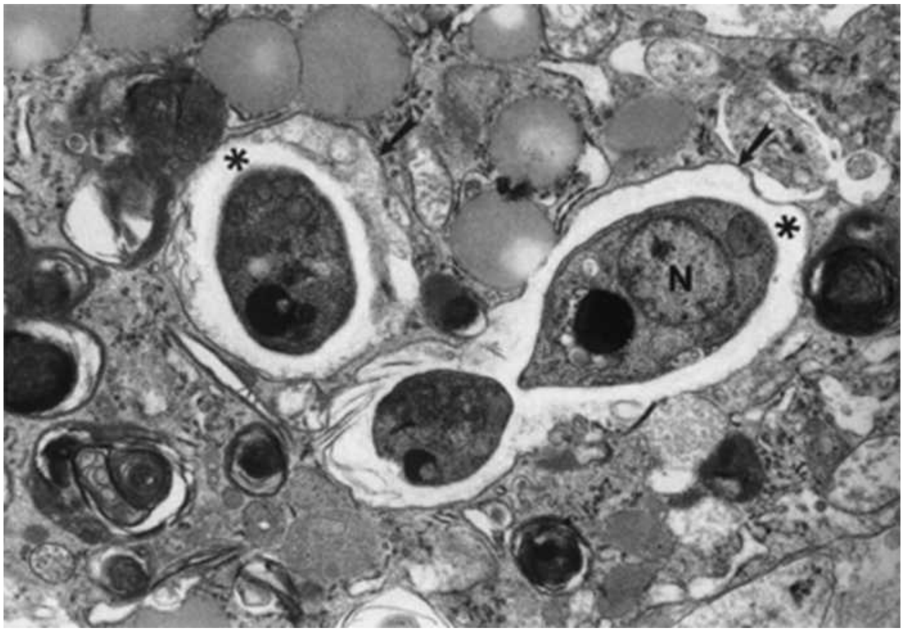


Figure 9. *Histoplasma capsulatum* (supraclavicular lymph node). High magnification of parasitic yeast forms illustrates details of their internal structure. N = nucleus; * = clear, peripheral, cytoplasmic halo; arrows = parasitic cell membrane. (20,000×) (From Ref. 10. Reproduced by courtesy of Springer-Verlag GmbH.)

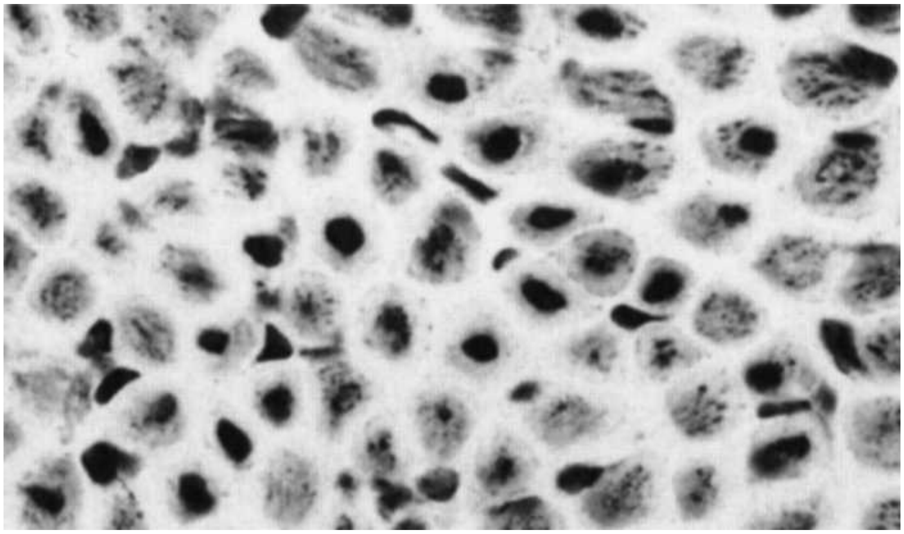


Figure 10. Centronuclear (myotubular) myopathy. The majority of small fibers contain centrally placed nuclei. (H&E, 150×) (From Ref. 10. Reproduced by courtesy of Springer-Verlag GmbH.)

electron microscope to reemerge as a valuable tool for analyzing molecular complexes.

Electron microscopy as an imaging technique allows a direct view of biological objects, while some of the other available techniques are indirect and in some instances nonspecific. Using electron microscopy, all components of the object and their mutual relationships at the molecular level can be analyzed. This information provides insight toward an understanding of structure–function relations.

The possibility of 3D reconstruction of cellular components via electron microscopy, along with the ease and speed with which newer instruments can provide data, have given the way to what many in the field are referring as a revolution. Recent technological advancements have made automated data acquisition possible, and have thus allowed a reduction of the total electron dose needed to image a specimen. Specimen preparation advances, such as embedding biological specimens in vitreous ice, have enabled studies of the macromolecular organization of cells. Whole prokaryotic and small eukaryotic cells can be directly grown and hydrated frozen on electron microscopy grids. Examination of the naturally preserved cells delivers images of the cellular structures in their functional environment. Such so-called tomograms contain all available information about the spatial relationships of macromolecular structures within the cell. However, due to their poor S/N and the generally highly crowded nature of the cytoplasm, the interpretation of

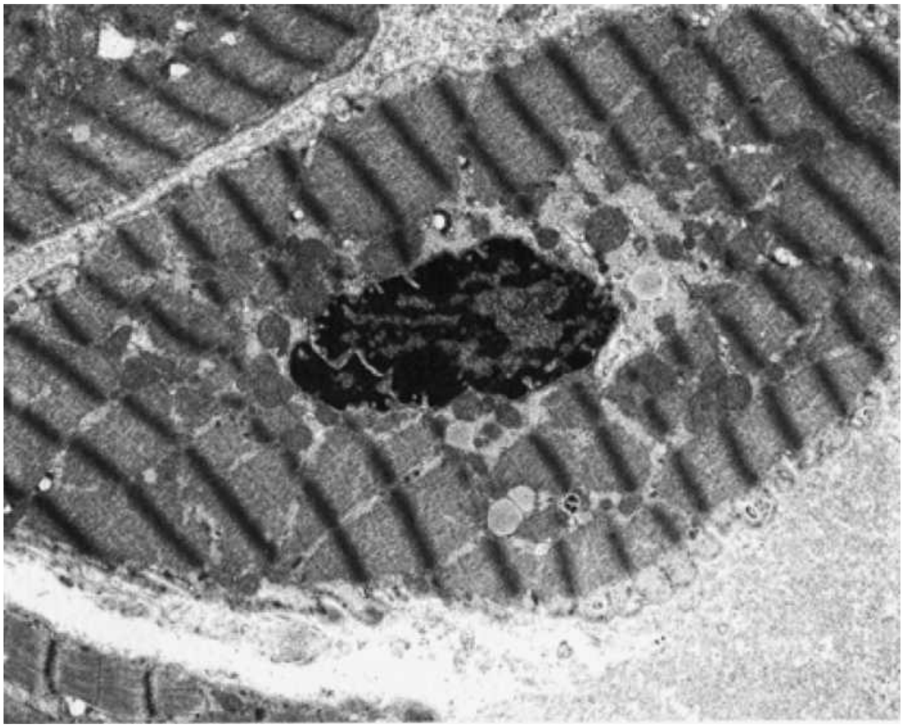


Figure 11. Muscle fiber with degenerated nucleus. Note absence of myofilaments and accumulation of glycogen in the paranuclear region to the right (15,000×). (From Ref. 10. Reproduced by courtesy of Springer-Verlag GmbH.)

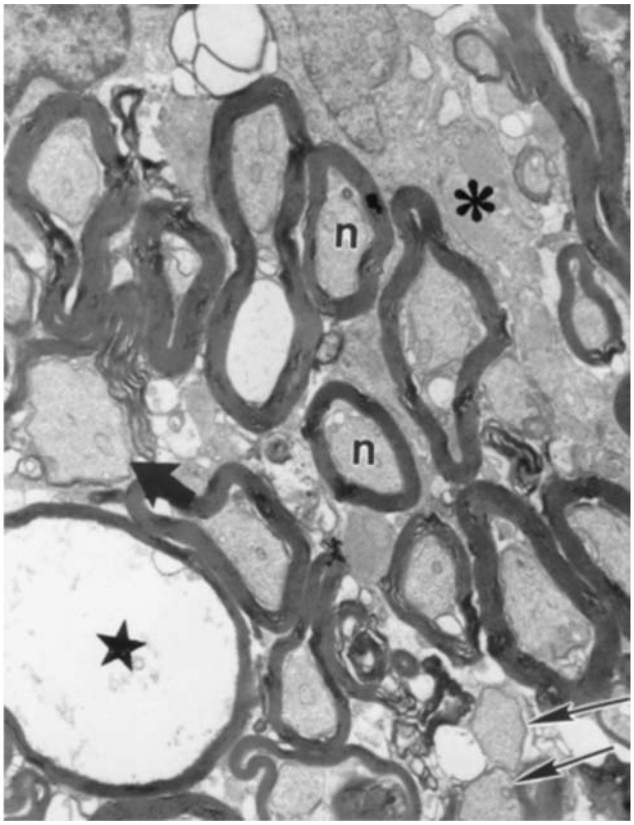


Figure 12. Ultrathin sections of opossum's optic nerve fibers 24 h after crash. Normal fibers (n) are seen among some altered fibers, which exhibit watery degeneration (star) and myelin sheath breakdown (thick arrow). Note demyelinated fibers (thin arrows) with an apparently intact axoplasmic cytoskeleton. Asterisk, astrocytic processes. (From Ref. 17. Reproduced by courtesy of Anais da Academia Brasileira de Ciencias.)

these tomograms remains difficult. To get significant information about specific structures in the cell, the images have to be evaluated using advanced pattern recognition methods. Existing structural models of cellular constituents at lower resolutions can guide the systematic evaluation of the tomograms. The aim is to visualize the complete 3D organization of the cell at molecular resolution. Structural evaluation by single particle analysis, electron crystallography and electron tomography is slow compared to other structure determination technologies, in particular X-ray crystallography. Processing time for the electronic technique is typically in the range of several months per solved structure, depending on the resolution achieved. The same task can be accomplished in the range of hours or days for X-ray crystallography, once suitable crystals are available. Continued joint efforts between the research community and manufacturers to develop user-friendly, universal interfaces between electron crystallography, single-particle analysis and electron tomography would improve this situation, and further expand the usefulness of of these electronic technologies.

BIBLIOGRAPHY

1. Nellist PD, et al Direct sub-angstrom imaging of a crystal lattice. Science 2004;305(5691):1741.
2. Freeman MR. Time -resolved scanning tunneling microscopy through tunnel distance modulation. Appl Phys Lett 1993;68(19):2633–2635.
3. Bonetta L. Zooming in on electron tomography. Nature Methods 2005;2(2):139–44.
4. Bozzola JJ, Russell LD. Electron Microscopy: Principals and techniques for biologists. Sudbury (MA): Jones and Bartlett Publishers; 1998.
5. Hayat MA. Principels and techniques of electron microscopy: Biological application. New York: Van Nostrand Reinhold Company; 1973.
6. Wischnitzer S. Introduction to electron microscopy. New York: Pergamon Press; 1981.
7. Meek GA. Practical electron microscopy for biologists. New York: John Wiley & Sons inc; 1976.
8. Joy DC. Beam interactions, contrast and resolution in the SEM. J Microsc 1984;136:241–58.
9. Haine M. The electron microscope: The present state of the Art. London: Spon; 1961.
10. Dickersin GR. Diagnostic Electron Microscopy: A text/ atlas. New York: Springer-Verlag; 1999.
11. Franchina M, Del Borrello E, Caruso A, Altavilla G. Serous tumors of the ovary: Ultrastructural observations. Eur J Gynaecol Oncol 1992;13(3):268–76.
12. Wolf HK, Garcia JA, Bossen EH. Oncocytic differentiation in intrahepatic biliary cystadenocarcinoma. Modern Pathol 1992;5(6):665–866.
13. Kobayashi TK, et al. Effects of Taxol on ascites cytology from a patient with fallopian tube carcinoma: Report of a case with ultrastructural studies. Diagn Cytopathol 2002;27(2):132–134.
14. Yogi T, et al Whipple's disease: The first Japanese case diagnosed by electron microscopy and polymerase chain reaction. Intern Med 2004;43(7):566–570.
15. Jensen HL, Norrild B. Herpes simplex virus-cell interactions studied by immunogold cryosection electron microscopy. Methods Mol Biol 2005;292:143–160.
16. Garrison RG, Boyd KS. Electron microscopy of yeastlike cell development from the microconidium of Histoplasma capsulatum. J Bacteriol 1978;133(1):345–353.
17. Narciso MS, Hokoc JN, Martinez AM. Watery and dark axons in Wallerian degeneration of the opossum's optic nerve: Different patterns of cytoskeletal breakdown? An Acad Bras Cienc 2001;73(2):231–243.

See also ANALYTICAL METHODS, AUTOMATED; CELLULAR IMAGING; CYTOLOGY, AUTOMATED.

MICROSCOPY, FLUORESCENCE

SERGE PELET
MICHAEL PREVITE
PETER T. C. SO
Massachusetts Institute of Technology
Cambridge, Massachusetts

INTRODUCTION

Fluorescence microscopy quantifies the distribution of fluorophores and their biochemical environment on the micron length scale and allows *In vivo* measurement of biological structures and functions (1–3). Heimstädt developed one of the earliest fluorescence microscopes in 1911. Some of the first biochemical applications of this technique include the study of living cells by the protozoologist Provazek in 1914.

Fluorescence microscopy is one of the most ubiquitous tools in biomedical laboratories. Fluorescence microscopy has three unique strengths. First, the fluorescence microscope has high biological specificity. Based on endogenous fluorophores or exogenous probes, fluorescence microscopy allows the association of a fluorescence signal with a specimen structural and biochemical state. While fluorescence microscopy has comparable resolution to white light microscopes, their range of applications in biomedicine is much broader.

Second fluorescence microscopy is highly sensitive in the imaging of cells and tissues. The high sensitivity of fluorescence microscopy originates from two factors. One factor is the significant separation between the fluorophores' excitation and emission spectra. This separation allows the fluorescence signal to be detected by efficiently rejecting the excitation radiation background using bandpass filters. The fluorescence microscope has the sensitivity to image even a single fluorophore. The other factor is the weak endogenous fluorescence background in typical biological systems. Since there is minimal background fluorescence, weak fluorescence signal from even a few fluorescent exogenous labels can be readily observed.

Third, fluorescence microscopy is a minimally invasive imaging technique. *In vivo* labeling and imaging procedures are well developed. While photodamage may still result from prolonged exposure of shorter excitation radiation, long-term observation of biological processes is possible. Today, a single neuron in the brain of a small animal can be imaged repeatedly over a period of months with no notable damage.

SPECTROSCOPIC PRINCIPLES OF FLUORESCENCE MICROSCOPY

Fluorescence Spectroscopy

An understanding of spectroscopic principles is essential to master fluorescence microscopy (4–6). Fluorescence is a photon emission process that occurs during molecular relaxation from electronic excited states. Historically, Brewster first witnessed the phenomenon of fluorescence in 1838 and Stokes coined the term fluorescence in 1852. These photonic processes involve transitions between electronic and vibrational states of polyatomic fluorescent molecules (fluorophores) by the absorption of either one or more photons. Electronic states are typically separated by energies on the order of $10{,}000\,\mathrm{cm}^{-1}$ and vibrational sublevels are separated by $\sim 10^2$–$10^3\,\mathrm{cm}^{-1}$. In a one-photon excitation process, photons with energies in the ultraviolet (UV) to the blue–green region of the spectrum are needed to trigger an electronic transition, whereas photons in the infrared (IR) spectral range are required for two-photon excitation. The molecules from the lowest vibrational level of the electronic ground state are excited to an accessible vibrational level in an electronic excited state. The molecule is quickly relaxed to the lowest vibrational level of the excited electronic state after excitation on the time scale of femtoseconds to picoseconds via vibrational processes. The energy loss in the vibrational relaxation process is the origin of the Stoke shift where fluorescence photons have longer wavelengths than the excitation radiation. The coupling of the ground and excited – state both for the absorption and emission process is governed by the Franck–Condon principle, which states that the probability of transition is proportional to the overlap of the initial and final vibrational wave function. Since the vibrational level structures of the excited and ground states are similar, the fluorescence emission spectrum is a mirror image of the absorption spectrum, but shifted to lower wavelengths. The shift between the maxima of the absorption and emission spectra is referred to as the Stokes' shift. The residence time of a fluorophore in the excited electronic state before returning to the ground state is called the fluorescence lifetime. The fluorescence lifetime is typically on the order of nanoseconds. The Jablonski diagram represents fluorescence excitation and deexcitation processes (Fig. 1).

Fluorescence deexcitation processes can occur via radiative and nonradiative pathways. Radiative decay describes molecular deexcitation processes accompanied by photon emission. Molecules in the excited electronic states can also relax by nonradiative processes where excitation energy is not converted into photons, but are dissipated by thermal processes, such as vibrational relaxation and collisional quenching. Let Γ and k be the radiative and nonradiative decay rates, respectively, and N be the number of fluorophore in the excited state. The temporal evolution of the excited state can be described by

$$\frac{dN}{dt} = -(\Gamma + k)N \tag{1}$$

$$N = N_0 e^{-(\Gamma + k)t} = N_0 e^{-t/\tau} \tag{2}$$

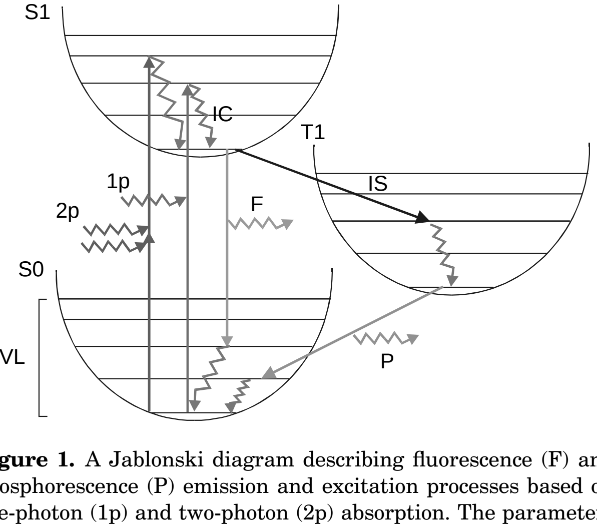


Figure 1. A Jablonski diagram describing fluorescence (F) and phosphorescence (P) emission and excitation processes based on one-photon (1p) and two-photon (2p) absorption. The parameters S0, S1, and T1 are the electronic singlet ground state, singlet excited state, and triplet excited state, respectively. Here VL denotes vibrational levels, IC denotes internal conversion, and IS denotes intersystem crossing.

The fluorescence lifetime, τ, of the fluorophore is the combined rate of the radiative and nonradiative pathways:

$$\tau = \frac{1}{\Gamma + k} \tag{3}$$

One can define the intrinsic lifetime of the fluorophore in the absence of nonradiative decay processes as, τ_0:

$$\tau_0 = \frac{1}{\Gamma} \tag{4}$$

The efficiency of the fluorophore can then be quantified by the fluorescence quantum yield, Q, which measures the fraction of excited fluorophore relaxing via the radiative pathway:

$$Q = \frac{\Gamma}{\Gamma + k} = \frac{\tau}{\tau_0} \tag{5}$$

Environmental Effect on Fluorescence

A number of factors contributes to the nonradiative decay pathways of the fluorophores and reduces fluorescence intensity. In general, the nonradiative decay processes can be classified as

$$k = k_{\text{ic}} + k_{\text{ec}} + k_{\text{et}} + k_{\text{is}} \tag{6}$$

where k_{ic} is the rate of internal conversion, k_{ec} is the rate of external conversion, k_{et} is the rate of energy transfer, and k_{is} is the rate of intersystem crossing.

Internal conversion describes the process where the electronic energy is converted to thermal energy via a vibrational process. The more interesting process is external conversion, where fluorophores lose electronic energy in collision process with other solutes. Several important solute molecules, such as oxygen, are efficient fluorescence quenchers. The external conversion process provides a convenient mean to measure the concentration of these molecules in the microenvironment of the fluorophore. The fluorophore is deexcited nonradiatively upon collision. The collisional quenching rate can be expressed as

$$k_{\text{ec}} = k_0[Q] \tag{7}$$

where $[Q]$ is the concentration of the quencher and k_0 is related to the diffusivity and the hydrodynamics radii of the reactants.

When collisional quenching is the dominant nonradiative process, equation 1 predicts that fluorescence lifetime decreases with quencher concentration.

$$\frac{\tau_0}{\tau} = (1 + k_0\tau_0[Q]) \tag{8}$$

Collision quenching also reduces the steady-state fluorescence intensity, F, relative to the fluorescence intensity in the absence of quencher, F_0. The Stern–Volmer equation describes this effect:

$$\frac{F_0}{F} = 1 + k_0\tau_0[Q] \tag{9}$$

A related process is steady-state quenching, where fluorescence signal reduction is due to ground-state processes. A fluorophore can be chemically bound to a quencher to form a dark complex, a product that does not fluoresce. In this case, steady-state fluorescence intensity also decreases with quencher concentration as

$$\frac{F_0}{F} = 1 + K_{\text{s}}[Q] \tag{10}$$

where K_{s} is the association constant of the quencher and the fluorophore. However, since steady-state quenching is a ground-state process that only reduces the fraction of fluorophores available for excitation, fluorescence lifetime is not affected.

Resonance energy-transfer rate, k_{et}, becomes significant when two fluorophores are in close proximity within ~5–10 nm as during molecular binding. The energy of an excited donor can be transferred to the accepted molecule via an induced dipole–induced dipole interaction. Let D represents the donor and A, the acceptor. Under illumination at the donor excitation wavelength, the number of excited donors and acceptors are N^{D}, N^{A}, respectively. Further, define the donor and acceptor deexcitation rates as k_{D} and k_{A}. The excited-state population dynamics of the donor and acceptor can be described as

$$\frac{dN^{\text{D}}}{dt} = -(k_{\text{D}} + k_{\text{et}})N^{\text{D}} \tag{11}$$

$$\frac{dN^{\text{A}}}{dt} = -k_{\text{A}}N^{\text{A}} + k_{\text{et}}N^{\text{D}} \tag{12}$$

Solving these equations provides the dynamics of donor and acceptor fluorescence:

$$N^{\text{D}} = N_0^{\text{D}}\exp[(-k_{\text{D}} - k_{\text{et}})t] \tag{13}$$

$$N^{\text{A}} = N_0^{\text{D}}\frac{k_{\text{et}}}{k_{\text{A}} - k_{\text{D}} - k_{\text{et}}}[\exp(-k_{\text{D}}t - k_{\text{et}}t) - \exp(-k_{\text{A}}t)] \tag{14}$$

The donor decay is a shortened single exponential, but the acceptor dynamics is more complex with two competing exponential processes.

The intersystem crossing rate, k_{is}, describes transitions between electronic excited states with wave functions of different symmetries. The normal ground state is a singlet state with an antisymmetric wave function. Excitation of the ground-state molecule via photon absorption results in the promotion of the molecule to an excited state with an antisymmetric wavefunction, another singlet state. Due to spin–orbit coupling, the excited molecule can transit into a triplet sate via intersystem crossing. The subsequent photon emission from the triplet state is called phosphorescence. Since the decay of the triplet state to the singlet ground state is forbidden radiatively, the triplet excited state has a very long lifetime on the order of microseconds to milliseconds.

FLUORESCENCE MICROSCOPE DESIGNS

The components common to most fluorescence microscopes are the light sources, the optical components, and the detection electronics. These components can be configured to create microscope designs with unique capabilities.

Fluorescence Excitation Light Sources

Fluorescence excitation light sources need to produce photons with sufficient energy and flux level. The ability to collimate the emitted rays from a light source further determines its applicability in high resolution imaging. Other less critical factors, such as wavelength selectivity, ease of use, and cost of operation, should also be considered.

Mercury arc lamps are one of the most commonly used light sources in fluorescence microscopy. The operation of a mercury arc lamp is based on the photoemission from mercury gas under electric discharge. The photoemission from a mercury arc consists of a broad background punctuated by strong emission lines. A mercury lamp can be considered as a quasimonochromatic light source by utilizing one of these strong emission lines. Since mercury lamps have emission lines throughout the near-UV and visible spectrum, the use of a mercury lamp allows easy matching of the excitation light spectrum with a given fluorophore by using an appropriate bandpass filter. Mercury arc lamps are also low cost and easy to use. However, since the emission of mercury lamps are difficult to collimate, they are rarely used in high resolution techniques, such as confocal microscopy. The advent of high power, energy efficient, light-emitting diodes (LEDs) with a long operation life allows the design of new light sources that are replacing arc lamps in some microscopy applications.

Laser light sources are commonly used in high resolution fluorescence microscopes. Laser light sources have a number of advantages including monochromatcity, high radiance, and low divergence. Due to basic laser physics, the laser emission is almost completely monochromatic. For fluorescence excitation, a monochromatic light source allows very easy separation of the excitation light from the emission signal. While the total energy emission from an arc lamp may be higher than some lasers, the energy within the excitation band is typically a small fraction of the total energy. In contrast, lasers have high radiance: the energy of a laser is focused within a single narrow spectral band. Therefore, the laser emission can be more efficiently used to trigger fluorescence excitation. Furthermore, laser emission has very low divergence and can be readily collimated to form a tight focus at the specimen permitting high resolution imaging. Gas lasers, such as the argon–ion laser and helium–neon lasers, are commonly used in fluorescence microscopy. Nowadays, they tend to be replaced by solid-state diode lasers that are more robust and fluctuate less. Lasers can further be characterized as continuous wave and pulsed. While continuous wave lasers are sufficient for most applications, pulsed lasers are used in two-photon microscopes where high intensity radiation is required for efficient induction of nonlinear optical effects.

Microscope Optical Components

The optical principle underlying fluorescence microscopes can be understood using basic ray tracing (7,8). The ray tracing of light through an ideal lens can be formulated into four rules: (1) A light ray originated from the focal point of a lens will emerge parallel to the optical axis after the lens. (2) A light ray propagating parallel to the optical axis will pass through the focal point after the lens. (3) Light rays originated from the focal plane of a lens will emerge collimated. (4) Collimated light rays incident upon a lens will focus at its focal plane (Fig. 2). From these rules, one can see that a simple microscope can be formed using two lenses with different focal lengths (Fig. 3). The lens, L1, with focal length, *f*1, images the sample plane and is called the objective. The lens, L2, with focal length, *f*2, projects the image onto the detector plane and is called the tube lens. From simple geometry, two points P1 and P2 separated by *x* in the sample plane will be separated by *x*(*f*2/*f*1) at the detector plane, where the ratio $M = f2/f1$ is called the magnification. One can see that the image in the sample plane is enlarged by the magnification factor at the detector.

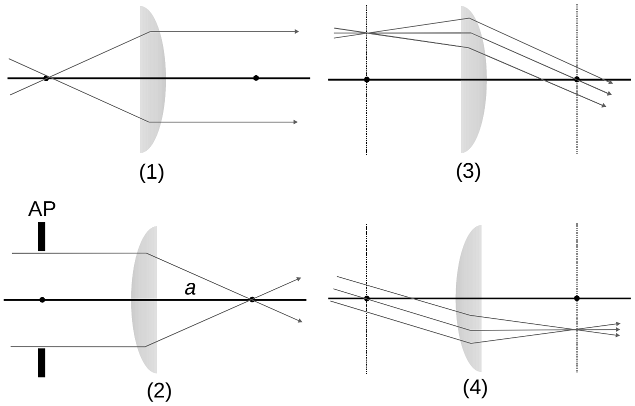


Figure 2. Four basic rules of optical ray tracing. (1) Light emerging from the focal point will become collimated parallel to the optical axis after the lens. And inversely (2) a collimated beam parallel to the optical axis will be focused at the focal plane of the lens. (3) A light source in the focal plane of the lens will become collimated after passing through the lens with an oblique angle determined by the distance from the optical axis and inversely (4) An oblique collimated beam will be focused in the focal plane by the lens. The numerical aperture of an imaging system is a function of the maximum convergence angle, α, as defined in rule 2. The maximum convergence angle is a function of the lens property and its aperture (AP) size.

By using the common wide-field fluorescence microscope as an example, we can further examine the components of a

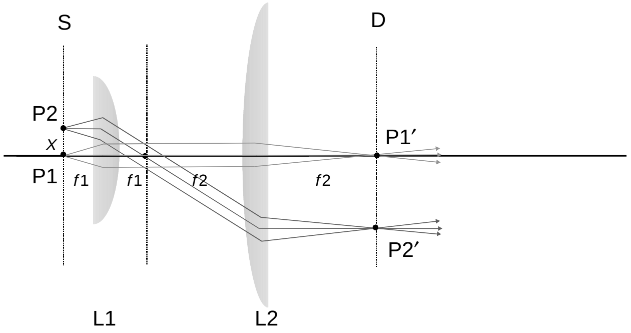


Figure 3. The detection path of a microscope. Lenses L1 and L2 are the objective and the tube lens, respectively. L1 has focal length *f*1 and L2 has focal length *f*2. For two points, P1 and P2 with separation, *x*, on the sample plane (S), these points are projected to points P1’ and P2’ on the detector plane (D).

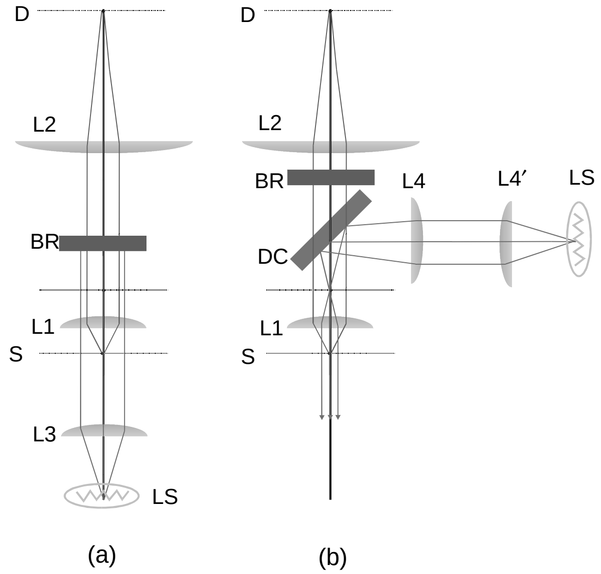


Figure 4. Two configurations of fluorescence microscopy (a) trans-illumination and (b) epi-illumination. The objective and detection tube lenses are L1 and L2. The condenser is L3. The excitation relay lenses are L4 and L4′. The sample and detector planes are S and D respectively. The light source is LS. The dichroic filter and the barrier filter are DC and BR.

complete fluorescence microscope system (Fig. 4a) (9). In addition to the detection optical path, fluorescence microscope requires an excitation light source. The excitation light source is typically placed in the focal point of a third lens, L3. The lens collimates the excitation light and projects it uniformly on the specimen (Koehler illumination). The lens, L3, is called the condenser. Since the excitation light is typically much stronger than the fluorescence emission, a bandpass filter is needed to block the excitation light. In this trans-illumination configuration, it is often difficult to select a bandpass filter with sufficient blocking power without also losing a significant portion of the fluorescence signal. To overcome this problem, an alternative geometry, epi-illumination, is commonly used (Fig. 4b). In this geometry, lens L1 functions both as the imaging objective and the condenser for the excitation light. A couple of relay lenses (L4, L4’) are used to focus the excitation light at the back aperture plane of the objective via a dichroic filter that reflects the excitation light but transmits the fluorescence signal. The excitation light is collimated by L1 and uniformly illuminates the sample plane. The fluorescence signal from the sample is collected by the objective and projected onto the detector via the tube lens L2. Since the excitation light is not directed at the detector, the task of rejecting excess excitation radiation at the detector is significantly easier. A barrier filter is still needed to eliminate stray excitation radiation from the optical surfaces.

From Fig. 2, one may assume that arbitrarily small objects can be imaged by increasing the magnification ratio. However, this is erroneous as the interference of light imposes a resolution limit on an optical system (10). The smallest scale features that can be resolved using fluorescence microscopy are prescribed by the Abbe limit. For an infinitely small emitter at the sample plane, the image at the detector, the point spread function (PSF), is not a single point. Instead, the intensity is distributed according to an Airy function with a diameter, d:

$$d = M\frac{1.22\lambda}{\mathrm{NA}} \tag{15}$$

where M is the magnification of the system, λ is the emission wavelength, and NA is the numerical aperture of the objective, which is defined as (Fig. 2):

$$NA = n \sin\alpha \tag{16}$$

where α is the half-convergence angle of the light and n is the index of refraction of the material between the lens and the sample. Therefore, the images of two objects on the sample plane will overlap if their separation is $<1.22\lambda/\mathrm{NA}$. Since NA is always on the order of 1, an optical system can only resolve two separate objects if their separation is on the order of the wavelength of light.

Fluorescence Detectors and Signal Processing

Since the fluorescence signal is relatively weak, sensitive detectors are crucial in the design of a high performance fluorescence microscope. For a wide field microscope, the most commonly used detectors are charged couple device (CCD) cameras, which are area detectors that contain a rectilinear array of pixels. Each pixel is a silicon semiconductor photosensor called a photodiode. When light is incident upon an individual photodiode, electrons are generated in the semiconductor matrix. Electrodes are organized in the CCD camera such that the charges generated by optical photons can be stored capacitatively during the data acquisition. After data acquisition, manipulating the voltages of the electrodes on the CCD chip allows the charges stored in each pixel to be extracted from the detector sequentially and read out by the signal conditioning circuit. These cameras are very efficient devices with a quantum efficiency up to ~80% (i.e., they can detect up to 8 out of 10 incident photons). Furthermore, CCD cameras can be made very low noise such that even four to five photons stored in a given pixel can be detected above the inherent electronic noise background of the readout electronics.

While CCDs are the detector of choice for wide-field microscopy imaging, there are other microscope configurations (discussed below) where an array detector is not necessary and significantly lower cost single element detectors can be used. Two commonly used single element detectors are avalanche photodiodes (APDs) and photomultiplier tubes (PMTs).

Avalanche photodiodes and photomultiplier tubes have been used in confocal and multiphoton microscopes. Avalanche photodiodes are similar to the photodiode element in a CCD chip. By placing a high voltage across the device, the photoelectron generated by the photon is accelerated across the active area of the semiconductor and collide with other electrons. Some of these electrons gain sufficient mobility from the collision and are accelerated toward the anode of the device themselves. This results in an avalanche effect with a total electron gain on the order of hundreds to thousands. A sizable photocurrent is generated for each input photon. A normal photodiode or a CCD camera does not have single photon sensitivity because the readout electronic noise is higher than the single electron level. The gain in the avalanche photodiode allows single photon detection. Photomultiplier tubes operate on a similar concept. A photomultiplier is not a solid-state device, but a vacuum tube where the photons impact the cathode and generates a photoelectron using the photoelectric effect. The electron generated is accelerated by a high voltage toward a second electrode, called a dynode. The impact of the first electron results in the generation of a cascade of new electrons that are then accelerated toward the next dynode. A photomultiplier typically has ~5–10 dynode stages. The electron current generated is collected by the last electrode, the anode, and is extracted. The electron gain of a photomultiplier is typically >1–10 million. While APDs and PMTs are similar devices, they do have some fundamental differences. The APD are silicon devices and have a very high quantum efficiency (~ 80%) from the visible to the near-IR spectral range. The PMT photocathode material has a typical efficient of 20%, but can reach ~ 40% in the blue–green spectral range. However, PMTs are not sensitive in the red–IR range with quantum efficiency dropping to a few percent. On the other hand, PMT have significantly higher gain and better temporal resolution.

Advanced Fluorescence Microscopy Configurations

In addition to wide-field imaging, fluorescence microscopy can be implemented in other more advanced configurations to enable novel imaging modes. We will cover four other particularly important configurations: wide-field deconvolution microscopy, confocal microscopy, two-photon microscopy, and total internal reflection microscopy.

Wide-Field Deconvolution Microscopy. Wide-field microscopy is a versatile, low cost, and widely used technique. However, cells and tissues are inherently three dimensional (3D). In a thick sample, the signals from multiple sample planes are integrated to form the final image. Since there is little correlation between the structures at different depths, the final image becomes fuzzy. The need for 3D resolved imaging has long been recognized. The iterative deconvolution approach has worked well for relatively thin specimen, such as in the imaging of organelle structures in cultured cells (11) (Fig. 5). In terms of instrument modifications, the main difference between deconvolution microscopy and wide-field microscope is the incorporation of an automated axial scanning stage allowing a 3D image stack to be acquired from the specimen. An initial estimate of the 3D distribution of fluorophores is convoluted with the known PSF of the optical system. The resultant image is then compared with the measured 3D experimental data. The differences allow a better guess of the actual fluorophore distribution. This modified fluorophore distribution is then convoluted with the system PSF again and allows another comparison with experimental data. This process repeats until an acceptable difference between the convoluted image and the experimental data is achieved. The deconvolution process in a wide-field fluorescence microscope belongs to the class of mathematical problems called ill-posed problems (12–14). An ill-posed problem does not have a unique solution, but depends on the selection of approach constraints to reach a final solution. One should consider the deconvoluted images only as the best estimate of the real physical structure given the available data. Furthermore, deconvolution algorithm is computationally intensive and often fails in thick specimens.

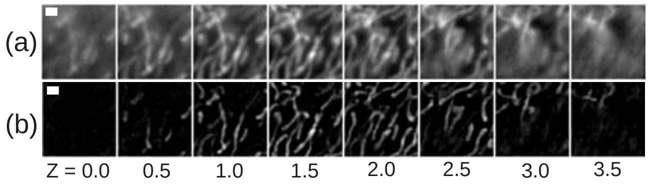


Figure 5. A comparison between (a) normal wide-field images and (b) deconvoluted images (11). Green fluorescent protein labeled mitochondria of a cultured cell was imaged by a wide-field fluorescence microscope as a 3D image stack. The image stack is deconvoluted and the significantly improved result is shown. The axial position of the image stack is shown below in units of micrometers.

Confocal Fluorescence Microscopy. Confocal fluorescence microscopy is a powerful method that can obtain 3D resolved sections in thick specimens by completely optical means (15–18). The operation principle of confocal microscopy is relatively straightforward. Consider the following confocal optical system in the transillumination geometry (Fig. 6). Excitation light is first focused at an excitation pinhole aperture. An excitation tube lens collimates the rays and projects them toward the condenser. The excitation light is focused at the specimen. The emitted light from the focal point is collected by the objective and collimated by the emission tube lens. The collimated light is subsequently refocused at the emission pinhole aperture. The detector is placed behind the aperture. As it is clear in the ray tracing illustration, the fluorescence signal produced at the specimen position defined by the excitation

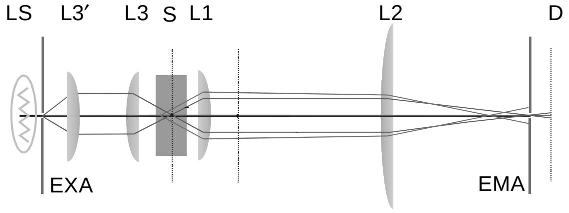


Figure 6. The configuration of a simple confocal microscope. The objective and the detection tube lenses are L1, L2. The light source is LS. The excitation aperture placed in front of the light source is EXA. The relay lens that images the excitation aperture and projects the image of the pinhole onto the specimen (S) are L3 and L3′. The fluorescence emission from the focal point (red rays) are projected onto the emission aperture (EMA) by L1 and L2. The signal is transmitted through EMA and is detected by the detector (D). Fluorescence generated outside the focal plane in the specimen (blue rays) are defocused at EMA and are mostly blocked.

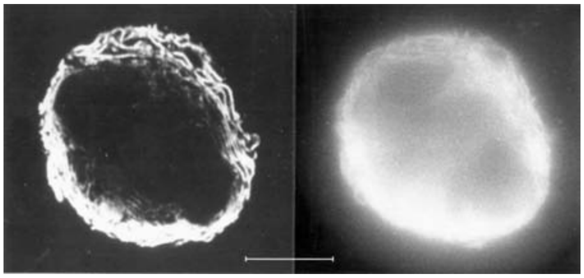


Figure 7. A comparison between confocal (a) and wide field (b) imaging of a plasmacytoma cell labeled with fluorescent antiendoplasmin that binds mainly to the endoplasmic reticulum. In the wide-field image, it is not possible to determine whether the central nucleic region contains endoplasmin end the structure of the cisternae are unclear (19).

pinhole aperture is exactly transmitted through the conjugate pinhole in the emission light path. However, for a fluorescence signal generated above or below the focal plane, the light is defocused at the emission pinhole aperture and is largely rejected. Hence, a pair of conjugate pinholes allows the selection of a 3D defined volume. One can show that a confocal microscope can image structures in 3D with a volume resolution of 0.1 fl. This system achieves 3D resolution, at the cost of obtaining fluorescence signal from only a single point in the specimen. It is necessary to raster scan the excitation focus to cover a 3D volume. Confocal microscopy has been used extensively to investigate microstructures in cells and in the imaging of tissues (19) (Fig. 7).

Two-Photon Fluorescence Microscopy. A two-photon microscope is an alternative to confocal microscopy for the 3D imaging of thick specimens. Denk, Webb, and coworkers in 1990 introduced two-photon excitation microscopy (18,20). Fluorophores can be excited by the simultaneous absorption of two photons each having one-half of the energy needed for the excitation transition. Since the two-photon excitation probability is significantly less than the one-photon probability, two-photon excitation occurs only at appreciable rates in regions of high temporal and spatial photon concentration. The high spatial concentration of photons can be achieved by focusing the laser beam with a high numerical aperture objective to a diffraction-limited spot. The high temporal concentration of photons is made possible by the availability of high peak power pulsed lasers (Fig. 8). Depth discrimination is the most important feature of multiphoton microscopy. In the two-photon case, >80% of the total fluorescence intensity comes from a 1 μm thick region about the focal point for objectives with numerical aperture of 1.25. For a 1.25 NA objective using excitation wavelength of 960 nm, the typical point spread function has a fwhm of 0.3 μm in the radial direction and 0.9 μm in the axial direction (Fig. 8). Two-photon microscopy has a number of advantages compared with confocal imaging: (*1*) Since a two-photon microscope obtains 3D resolution by limitation the region of excitation instead of the region of detection as in a confocal system, photodamage of biological specimens is restricted to the focal point. Since out-of-plane chromophores are not excited, they are not subject to photobleaching. (*2*) Two-photon excitation wavelengths are typically redshifted to about twice the one-photon excitation wavelengths in the IR spectral range. The absorption and scattering of the excitation light in thick biological specimens are reduced. (*3*) The wide separation between the excitation and emission spectra ensures that the excitation light and Raman scattering can be rejected without filtering out any of the fluorescence photons. An excellent demonstration of the ability of two-photon imaging for deep tissue imaging is in the neurobiology area (21) (Fig. 9).

Total internal reflection microscopy. Confocal and two-photon microscopy can obtain 3D resolved images from specimens up to a few hundred micrometers in thickness. However, both types of microscopy are technically challenging, require expensive instrumentation, and only can acquire data sequentially from single points. Total internal reflection microscopy (TIRM) is an interesting alternative if 3D-resolved information is only required at the bottom surface of the specimen, such as the basal membrane of a cell (22–24). Total internal reflection occurs at an interface between materials with distinct indices of refraction (Fig. 10). If light ray is incident from a high index prism, n_2, toward the lower index region, n_1, at an angle θ, the light will be completely reflected at the interface if θ is $>\theta_c$, the critical angle.

$$\sin\theta_c = \frac{n_1}{n_2} \tag{17}$$

While the light is completely reflected at the interface, the electric field intensity right above the interface is nonzero, but decays exponentially into the low index medium. The decay length of the electric field is on the order of tens to hundreds of nanometers. Compared with other forms of 3D resolved microscopy, TIRM allows the selection of the thinnest optical section, but only at the lower surface of the sample. While prism launch TIRM as described is simpler to construct, the bulky prism complicates the routine use of TIRM for cell biology studies. Instead, ultrahigh numerical aperture objectives have been produced (1.45–1.6 N). Light rays focus at the back aperture plane of the objective that are sufficiently off axis will emerge collimated, but at an oblique angle. If a specimen grown on a high index coverglass is placed upon the objective, total internal reflection can occur at the specimen-coverglass interface if the oblique angle is sufficiently large. This approach has been described as the objective launch TIRM and has been very successful in the study of exocytosis processes (23) (Fig. 11).

FLUORESCENT PROBES

Fluorescence microscopy has found many applications in biomedicine. This wide acceptance is a direct result of the availability of an ever growing set of fluorescence probes designed to measure cell and tissue structure, metabolism, signaling processes, gene expression, and protein distribution (25,26). The synthesis of fluorescent probes dates back to 1856, when William Perkin made the first synthetic

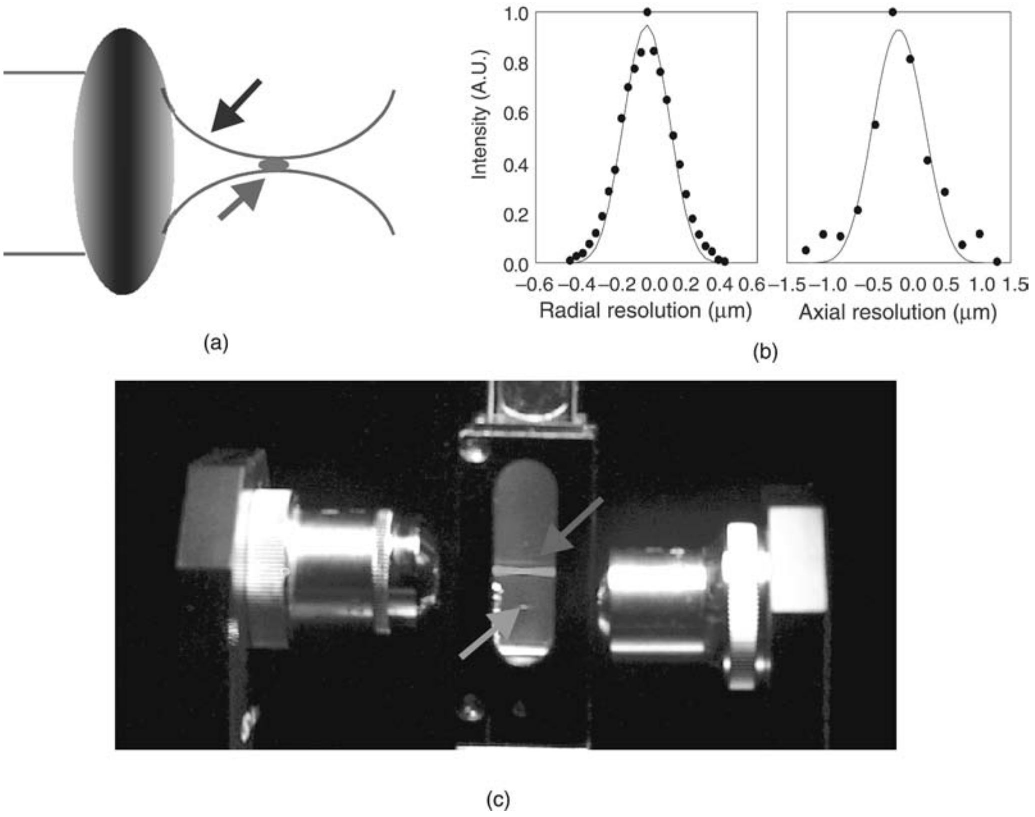


Figure 8. Two-photon microscopy optically sections and produces a fluorescent signal originating only from the focal point (a) the geometry of two-photon fluorescence. In traditional one-photon excitation, fluorescence is generated throughout the double inverted cones (blue arrow). Two-photon excitation generates fluorescence only at the focal point (red arrow). (b) The submicron PSF of two-photon excitation at 960 nm: The full-widths at half maximum (fwhm) are 0.3 μm radially and 0.9 μm axially. (c) An experimental visualization of the small excitation volume of two-photon fluorescence. One- and two-photon excitation beams are focused by two objectives (equal numerical aperture) onto a fluorescein solution. Fluorescence is generated all along the path in the one-photon excitation case (blue arrow), whereas fluorescence is generated only in a 3D confined focal spot for two-photon excitation (red arrow) The reduced excitation volume is thought to lead to less photodamage. (Please see online version for color figure)

probe from coal tar dye. Thereafter, many more synthetic dyes became available: pararosaniline, methyl violet, malachite green, safranin O, methylene blue, and numerous azo dyes. While most of these early dyes are weakly fluorescent, more fluorescent ones based on the xanthene and acridine heterocyclic ring systems soon became available.

Optical Factors in the Selection of Fluorescent Probes

Before providing a survey of the wide variety of fluorescent probes, it is important to first discuss the optical properties of fluorescent probes that are important for microscopic imaging: extinction coefficient, quantum yield, fluorescent lifetime, photobleaching rate, and spectral characteristics.

One of the most important characteristic of a fluorescent probe is its extinction coefficient. Extinction coefficient, ϵ, measures the absorption probability of the excitation light by the fluorophore. Consider excitation light is transmitted through a solution containing fluorophore at concentration c with a path length l. The light intensities before and after the solution are I_0 and I. The extinction coefficient can then be defined by Beer's law:

$$\log_{10}\frac{I_0}{I} = \varepsilon cl \tag{18}$$

Fluorescent probes with high extinction coefficients can be excited by lower incident light intensity allowing the use of lowest cost light sources and reducing the background noise of the images originated from scattered excitation light.

Quantum yield, Q, measures the relative contributions of the radiative versus nonradiative decay pathways. High quantum efficiency maximizes the fluorescent signal for each photon absorbed. The combination of probe extinction coefficient and quantum efficiency quantifies the total conversion efficiency of excitation light into fluorescent signal.

While ϵ and Q determines excitation light conversion efficiency, the maximum rate of fluorescent photon generation also depends on the lifetime, τ, of the probe. Since a molecule that has been excited cannot be reexcited until it returns to the ground state, fluorescent lifetime

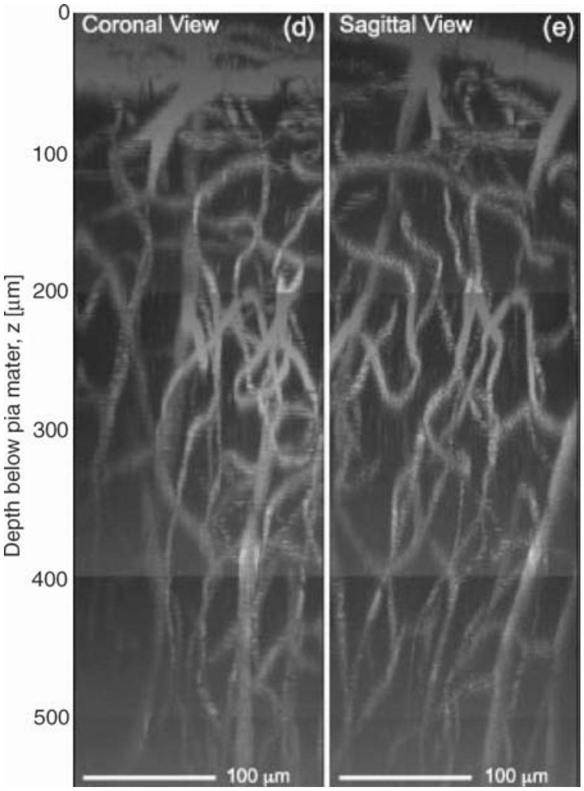


Figure 9. Fluorescein dextran labeled blood vessels in the primary vibrissa cortex of a living rat brain imaged using two-photon microscope down to a depth of $>500\,\mu$m, which demonstrates the ability of this technique to image deep into tissue (21).

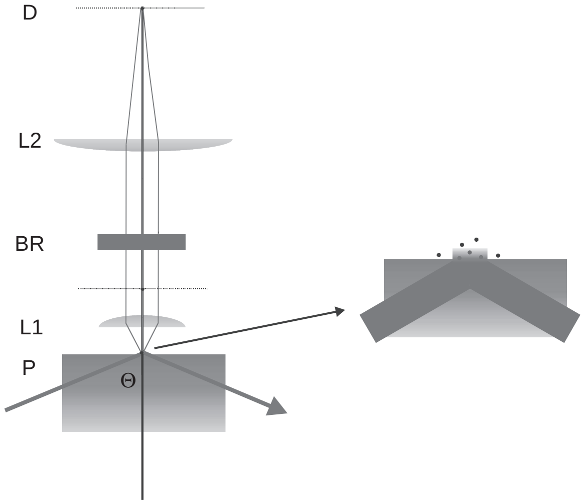


Figure 10. The configuration of a total internal reflection fluorescence microscope. L1 and L2 are objective and tube lens, respectively. The barrier filter is BR and the detector is D. The prism is P. The excitation light (green) is incident up the prism at angle, θ, greater than the critical angle. The excitation light is totally internally reflected from the surface. A magnified view is shown on the left. The evanescence electric field induced by the excitation light above the prism surface decays exponentially and only induces strong fluorescence signal for probes close to the surface of the prism. Please see online version for color figure.

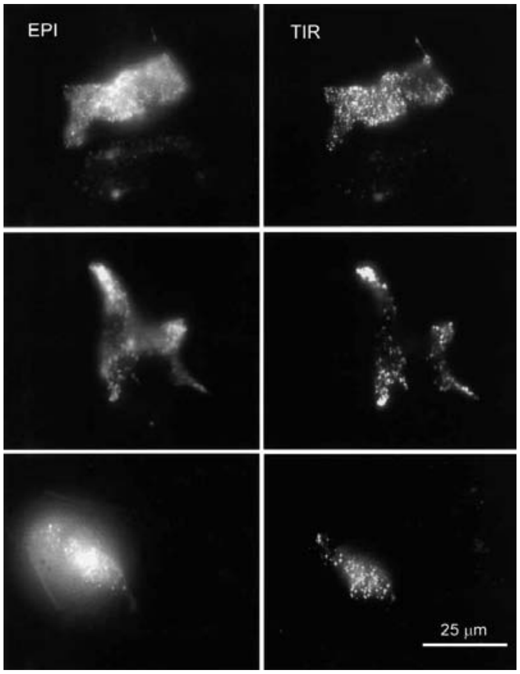


Figure 11. Epi-illuminated wide field (EPI) and total internal reflection (TIR) microscopy of bovine chromaffin cells containing secretory granules marked with GFP atrial naturetic protein (23). Only the lower plane of the cells contributes to the fluorescence signal in TIR set-up.

determines the rate at which a single probe molecule can be recycled. In general, for fluorescent probes with equal ε and Q, fluorescent photon production rate is an inverse function of probe lifetime. Further intersystem cross-rate also plays a role in determining photon generation rate. Since the triplet state has a very long lifetime, probes with high intersystem cross-rates are trapped in the triplet state with a relatively lower photon generation rate.

Photobleaching rate measures the probability that a probe will undergo an excited-state chemical reaction and become nonfluorescent irreversibly. Therefore, the photobleaching rate of a probe limits the maximum number of fluorescence photons that can be produced by a single fluorophore. Photobleaching rates of fluorophores vary greatly. For example, rhodamine can survive up to 100,000 excitation, fluorescein a few thousand, and tryptophan can only sustain a few excitation events. Photobleaching can be caused by a variety of processes. Generally, it is the result of a photochemical reaction in the excited state of the probe. For example, a common bleaching pathway is the generation of a triplet state that reacts with oxygen dissolved in solution to generate singlet oxygen and an oxidized molecule incapable of undergoing the same electronic transition as before.

Spectral properties are also important in probe selection for a number of reasons. First, selecting fluorescent probes with well-separated excitation and emission spectra allow more efficient separation of the fluorescence signal from the excitation light background. Second, fluorescent probes should be selected to match the detector used in the microscope that may have very different sensitivity across the spectral range. For example, most PMTs have maximum efficiency in the green spectral range, but very low efficiency in the red. Therefore, green emitting probes are often better matches for microscopes using PMTs as detectors. Third, probes with narrow emission spectra allow a specimen to be simultaneously labeled with different colors providing a method to analyze multiple biochemical components simultaneously in the specimen.

Classification of Fluorescent Probes

There is no completely concise and definitive way to classify the great variety of fluorescent probes. A classification can be made based on how the fluorophores are deployed in biomedical microscopy: intrinsic probes, extrinsic probes, and genetic expressible probes.

Intrinsic Probes. Intrinsic probes refer to the class of endogenous fluorophores found in cells and tissues. Many biological components, deoxyribonuclic acid such as (DNA), proteins, and lipid membrane are weakly fluorescent. For example, protein fluorescence is due to the presence of amino acids: tryptophan, tyrosine, and phenylalanine. Among them, tryptophan is the only member with marginal quantum yield for microscopic imaging. However, fluorescent imaging based on tryptophan provides very limited information due to the prevalence of this amino acid in many proteins distributed throughout cellular systems and provides no specificity or contrast. The most useful intrinsic probes for microscopy imaging are a number of enzymes and proteins, such as reduced pyridine nucleotides [NAD(P)H], flavoproteins, and protoporphyrin IX. Both NAD(P)H and favoproteins are present in the cellular redox pathway. The NAD(P)H becomes highly fluorescent when reduced, whereas flavoprotein becomes fluorescent when oxidized, while their redox counterparts are nonfluorescent. These enzymes thus provide a powerful method to monitor cell and tissue metabolism. Protoporphyrin IX (PPIX) is a natural byproduct in the heme production pathway that is highly fluorescent. Certain cancer cells have been shown to have upregulate PPIX production relative to normal tissue and may be useful in the optical detection of cancer. Another class of important intrinsic fluorophores includes elastin and collagen, which resides in the extracellular matrix allowing structural imaging of tissues. Finally, natural pigment molecules, such as lipofuscin and melanin, are also fluorescent and have been used in assaying aging in the ocular system and malignancy in the dermal system respectively.

Extrinsic Probes. Many extrinsic fluorescent probes have been created over the last century. A majority of these extrinsic fluorophores are small aromatic organic molecules (25–28). Many probe families, such as xanthenes, canines, Alexas, coumarines, and acrinides have been created. These probes are designed to span the emission spectrum from near UV to the near-IR range with optimized optical properties. Since these molecules have no intrinsic biological activity, they must be conjugated to biological molecules of interest, which may be proteins or structure components, such as lipid molecules. Most common linkages are through reactions to amine and thiol residues. Reactions to amine are based on acylating reactions to form carboxamides, sulfonamides, or thioureas. Targeting thiol residue in the cysteines of proteins can be accomplished via iodoacetamides or maleimides. Other approaches to conjugate fluorophores to biological components may be based on general purpose linker molecules, such as biotin-avidin pairs or based on photoactivable linkers. A particularly important class of fluorophores conjugated proteins is fluorescent antibodies that allow biologically specific labeling.

In addition to maximizing the fluorescent signal, the greater challenge in the design of small molecular probes is to provide environmental sensitivity. An important class of environmentally sensitive probes distinguishes the hydrophilic versus hydrophobic environment and results in a significant quantum yield change or spectral shift based on solvent interaction. This class of probes includes DAPI, laurdan, and ANS, which have been used to specifically label DNA, measure membrane fluidity, and sense protein folding states, respectively. Another important class of environmentally sensitive probes senses intracellular ion concentrations, such as pH, Ca^{2+}, Mg^{2+}, Zn^{2+}. The most important members of this class of probes are calcium concentration sensitive because of the importance of calcium as a secondary messenger. Changes in intracellular calcium levels have been measured by using probes that either show an intensity or a spectral response upon calcium binding. These probes are predominantly analogues of calcium chelators. Members of the Fluo-3 series and Rhod-2 series allow fast measurement of the calcium level based upon intensity changes. More quantitative measurement can be based on the Fura-1 and Indo-1 series that are ratiometric. These probes exhibit a shift in the excitation or emission spectrum with the formation of isosbestic points upon calcium binding. The intensity ratio between the emission maxima and the isosbestic point allows a quantitative measurement of calcium concentration without influence from the differential partitioning of the dyes into cells.

Quantum dots belong to a new group of extrinsic probes that are rapidly gaining acceptance for biomedical imaging due to a number of their very unique characteristics (29–31). Quantum dots are semiconductor nanoparticles in the size range of 2–6 nm. Photon absorption in the semiconductor results in the formation of an exciton (an electron-hole pair). Semiconductor nanoparticles with diameters below the Bohr radius exhibit strong quantum confinement effect, which results in the quantization of their electronic energy level. The quantization level is related to particle size where smaller particles have a larger energy gap. The radiative recombination of the exciton results in the emission of a fluorescence photon with energy corresponding to the exciton's quantized energy levels. The lifetime for the recombination of the exciton is long, typically on the order of a few tens of nanoseconds. Quantum dots have been fabricated from II–VI (e.g., as CdSe, CdTe, CdS, and ZnSe) and III–V (e.g., InP and InAs) semiconductors. Due to the compounds involved in the formation of these fluorescent labels, toxicity studies have to be realized prior to any experiments. Recent research works have been devoted to the better manufacture of these semiconductor crystals including methods to form a uniform crystalline core and to produce a surface capping layer that enhances the biocompability of these compounds, prevents their aggregation, and can maximize their quantum efficiency. Furthermore, coating the surface of quantum dots with convenient functional groups, including common linkages, such as silane or biotin, has been accomplished to facilitate linkage to the biological molecules. Quantum dots are unique in their broad absorption spectra, very narrow (~15 nm) emission spectrum, and extraordinary photostability. In fact, quantum dots have been shown to have photobleaching rates orders of magnitude below that of organic dyes. Quantum dots also have excellent extinction coefficients and quantum yield. While there are significant advantages in using quantum dots, they also have a number of limitations including their relative larger size compared with organic dyes and their lower fluorescence flux due to their long lifetime. Quantum dots have been applied for single receptor tracking on cell surface and for the visualization of tissue structures, such as blood vessels.

Genetic Expressible Probes. The development of genetically expressible probes has been rapid over the last decade (32). The most notable of these genetic probes is green fluorescent protein, GFP (33). The GFP was isolated and purified from the bioluminescent jellyfish *Aequorea Victoria*. Fusion proteins can be created by inserting GFP genes into an expression vector that carries a gene coding for a protein of interest. This provides a completely noninvasive procedure and perfectly molecular specific approach to track the expression, distribution, and trafficking of specific proteins in cells and tissues. In order to better understand protein signaling processes and protein–protein interactions, fluorescent proteins of different colors have been created based on random mutation processes. Today, fluorescent proteins with emission spanning the spectral range from blue to red are readily available. Expressible fluorescent proteins that are sensitive to cellular biochemical environment, such as pH and calcium, have also been developed. Novel fluorescent proteins with optically controllable fluorescent properties, such as photoactivatable fluorescent proteins, PA-GFP, photoswitchable CFP, and pKindling red have been created and may be used in tracing cell movement or protein transport. Finally, protein–protein interactions have been detected based on a novel fluorescent protein approach in which each of the interacting protein pairs carries one-half of a fluorescent protein structure that is not fluorescent. Upon binding of the protein pairs, the two halves of the fluorescent protein also recombine, which results in a fluorescent signal.

ADVANCED FUNCTIONAL IMAGING MODALITIES AND THEIR APPLICATIONS

A number of functional imaging modalities based on fluorescent microscopy have been developed. These techniques are extremely versatile and have found applications ranging from single molecular studies to tissue level experiments. The implementation of the most common imaging modalities will be discussed with representative examples from the literature.

Intensity Measurements

The most basic application of fluorescence microscopy consists in mapping fluorophore distribution based on their emission intensity as a function of position. However, this map is not static. Measuring intensity distribution as a function of time allows one to follow the evolution of biological processes. The fastest wide-field detectors can have a frame rate in the tens of kilohertz range, unfortunately at the expense of sensitivity and spatial resolution. They are used to study extremely fast dynamics, such as membrane potential imaging in neurons. For 3D imaging, point scanning techniques are typically slower than wide-field imaging, but can reach video rate speed using multifoci illumination.

Dynamic intensity imaging has been used at the tissue level to follow cancer cells as they flow through blood vessels and extravasate to form metastases, or in embryos to track the expression of a regulatory protein at different developmental stages. One commonly used technique to follow the movements of protein in cellular systems is fluorescent recovery after photobleaching (FRAP). In FRAP studies, a small area of a cell expressing a fluorescently labeled protein is subjected to an intense illumination that photobleaches the dye and leads to a drastic drop in fluorescence intensity. The rate at which the intensity recovers provides a measure of the mobility of the protein of interest.

An important aspect of the fluorescent microscopy technique lies also in the image analysis. Particle tracking experiments are an excellent example. Zuhang and coworkers (34) studied the infection pathway of the influenza virus labeled with a lipophilic dye in CHO cells. Each frame of the movie recorded was analyzed to extract the position of the virus particles with 40 nm accuracy. Three different stages of transport after endocytosis of the virus particle were separated, each involving different transport mechanisms transduced by a different protein as shown on Fig. 12. The first stage is dependant on actin and results in an average transport distance of 2 μm from the initial site of binding at the cell periphery. The second stage is characterized by a sudden directed displacement that brings the virus close to the nucleus with a speed of 1–4 $\mu\text{m s}^{-1}$ that is consistent with the velocity of dynein motors on microtublues. The last stage consists of back and forth motion in the perinuclear region. This is followed by the fusion of the endosome with the virus and the liberation of the genetic material. This event is identified by a sudden increase in the fluorescence intensity due to the dequenching of the fluorescent tags on the virus.

Spectral Measurements

An extremely important feature of fluorescent microscopy is the ability to image many different fluorescent species based on their distinct emission spectra. Dichroic bandpass filters optimized for the dyes used in the experiment can discriminate efficiently between up to four or five different fluorophores.

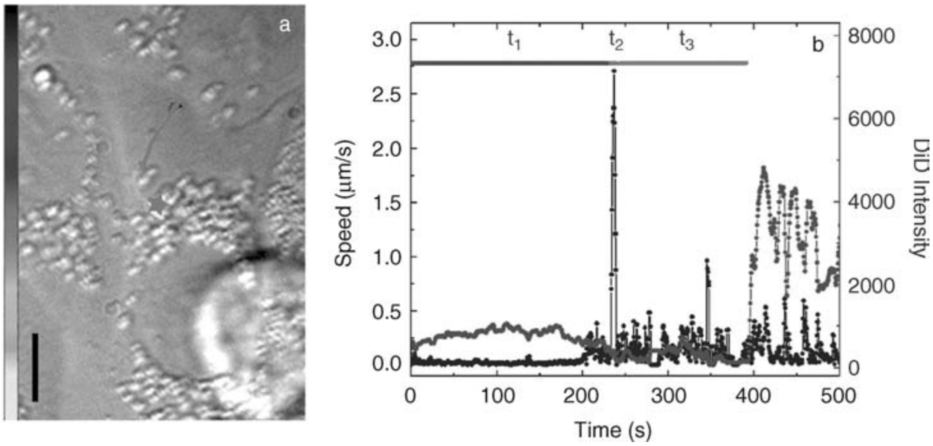


Figure 12. Particle tracking of virus infecting a cell. (a) Trajectory of the virus. The color of the trajectory codes time from 0 s (black) to 500 s (yellow). The star indicates the fusion site of the virus membrane with the vesicle. (b) Time trajectories of the velocity (black) and fluorescence (blue) of the virus particle (34). Please see online version for color figure.

In a study of connexin trafficking, Ellisman and coworkers (35) used a dual labeling scheme to highlight the dynamics of these proteins. Using a recombinant protein fused to a tetracystein receptor domain, the connexin was stably labeled with a biarsenical derivate of fluorescein or resorufin (a red fluorophore). The cells expressing these modified proteins were first stained with the green fluorophore and incubated 4–8 h. The proteins produced during this incubation period are fluorescently tagged in a second staining step with the red fluorophore. The two-color images highlight the dynamics of the connexin refurbishing at the gap junction. As shown on Fig. 13, the older proteins are found in the center and are surrounded by the newer proteins.

For wide-field fluorescence imaging using a CCD camera, spectral information is collected sequentially while position information is collected at once. Bandpass filters can be inserted to select the emission wavelength in between image frames. This procedure is relatively slow and can result in image misregistration due to the slight misalignment of the filters. This problem can be overcome by the use of electronically tunable filter. Two types of electronically tunable filters are available based either on liquid-crystal technology or on electrooptical crystals. Liquid-crystal tunable filters are made of stacks of birefringent liquid-crystal layers sandwiched between polarizing filters. Polarized light is incident upon the device. The application of a voltage on the liquid-crystal layer produces a wavelength dependent rotation of the polarization of light as the light is transmitted through the liquid-crystal layers. After cumulative rotations through the multiple layers, only the light at a specific spectral range is at the correct polarization to pass through the final polarizer without attenuation. The second type is acoustooptics tunable filters (AOTFs). An AOTF works by setting acoustic vibration at radio frequency (rf) through an electrooptical crystal to create a diffraction grating that singles out the appropriate wavelength with a few nanometer bandwidth. The main advantage of AOTF is that the wavelength selection is realized by tuning the acoustic wave frequency, which can be done in a fraction of a millisecond while the liquid-crystal tunable filters operate with a time constant of hundreds of milliseconds. The latter, however, have a larger clear aperture and selectable bandwidth ranging from a fraction of a nanometer up to tens of a nanometer. Liquid-crystal filters are more often used for emission filtering while the acoustooptic filters are more commonly used for excitation wavelength selection.

Typical emission spectra from molecular fluorophores have a sharp edge at the blue end of the spectrum, but have a long tail extending far into the red due to electronic relaxation from the excited state into a vibrationally excited ground state. When imaging with a few color channels, where each channel represents a single chromophore, one has to be careful to take into account the spectral bleedthrough of each dye into the neighboring channels. Collecting signal in a larger number of channels allows the use of a linear unmixing technique to account for the real shape of the emission spectra of each dye and accounts more precisely for their contributions in each pixel of the image. This technique can be implemented using tunable filters with a narrow bandwidth and CCD camera detectors. It has also been shown that an interferometer can be used to encode the spectral information in the image on the CCD camera. An image is then recorded for each step of the interferometer and a Fourier transform analysis allows the recovery of the spectral information. Although it requires more advanced postprocessing of the image data, this approach offers a large spectral range and a variable spectral resolution unmatched by the tunable filters.

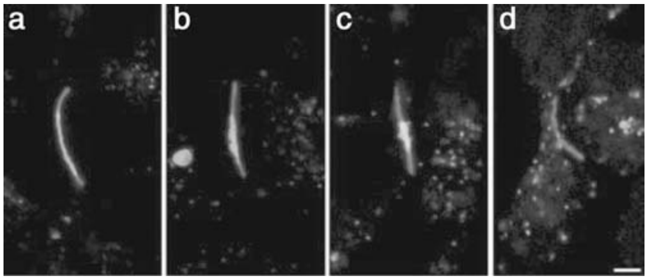


Figure 13. Connexin trafficking at gap junction. The newly produced proteins are labeled in red after 4 h (a and b) or 8 h (c and d) hours after the first staining step with green. The older proteins occupy the periphery of the gap junction, while the new ones are localized in its center (35). Please see online version for color figure.

In scanning systems, such as confocal microscopes, the use of dichroic beamsplitters can be readily constructed to simultaneously resolve two or three spectral channels in parallel at each scan position. If more spectral channels are desired for spectral decomposition measurement, the emission can be resolved in a multichannel detector using a grating or a prism to separate the different wavelength components. This has been used to separate the contribution of dyes with very similar emission spectra like GFP and fluorescein, or resolve the different intrinsic fluorophores contained in the skin where many fluorophores with overlapping spectra are present.

A particularly promising class of probes for spectral imaging are quantum dots. As discussed previously, the emission spectra of quantum dots are very narrow and can be tuned by changing their size. Further, all quantum dots have a very broad excitation spectrum and a single excitation wavelength larger than the band gap energy can lead to the efficient excitation of many different colored quantum dots simultaneously. In their report, Simon and coworkers (36) used these particles to track metastatic cells injected in the tail of a mouse as they extravasate into lung tissue. Using spectrally resolved measurements, they demonstrate their ability to recognize at least five different cell populations each labeled with different quantum dots. Figure 14 shows an image of cells labeled with different quantum dots and the emission spectra from each of these particles. The difference in emission spectra allows an easy identification of each cell population.

Lifetime Resolved Microscopy

Measurement of the fluorescence lifetime in a microscope provides another type of contrast mechanism and can be used to discriminate dyes emitting in the same wavelength range. It is also commonly used to monitor changes in the local environment of a probe measuring the pH or the concentration of cations *In situ*. The fluorescence lifetime can be shortened by interaction of the probe with a quencher, such as oxygen. Another type of quenching is induced by the presence of the transition dipole of other dyes, which are in close vicinity and lifetime measurements can be used to quantify energy-transfer processes (discussed further in a later section).

There are two methods to measure the fluorescence lifetime in a microscope. One is in the time domain and the other is in the frequency domain. In the time domain, a light pulse of short duration excites the sample and the decay of the emission is timed. The resulting intensity distribution is a convolution between the instrument response and the exponential decay of the fluorophore.

$$I(t) = I_0 \int_0^t G(t-T)\cdot \exp\left(\frac{T}{\tau}\right) dT \qquad (19)$$

In the frequency domain, the excitation light is modulated at frequency ω. The intrinsic response time of the fluorescence acts as a low pass filter and the emitted signal is phase shifted and demodulated. Both the demodulation and the phase shift can be linked to the fluorescence lifetime.

$$\Delta\phi = a\tan(\omega\tau) \qquad (20)$$

$$M = \frac{1}{\sqrt{1+\omega^2\tau^2}} \qquad (21)$$

In order to obtain a measurable phase shift and modulation, the frequency has to be on the same order of magnitude as the lifetime (i.e., 10^8 Hz). However, it is difficult to measure these two parameters at such high frequencies. Therefore, one typically uses a heterodyne detection to lower the frequency to the kilohertz range by modulating the detector at a frequency close to the excitation frequency.

For wide-field microscopy, an image intensifier is placed in front of the CCD camera to modulate the gain of detection. In the time domain, a short time gate is generated to collect the emission at various times after the excitation. In the frequency domain, the image intensifier is modulated at high frequencies and a series of images at different phases are acquired. In laser scanning confocal and multiphoton microscopes, time correlated single-photon counting is the method of choice for lifetime measurements in the time domain because it offers an excellent signal/noise ratio at low light levels. Every time a photon is detected, the time elapsed since the excitation of the sample is measured. A histogram of all the times of arrival yields a decay curve of the fluorescence in each pixel of the image. For brighter

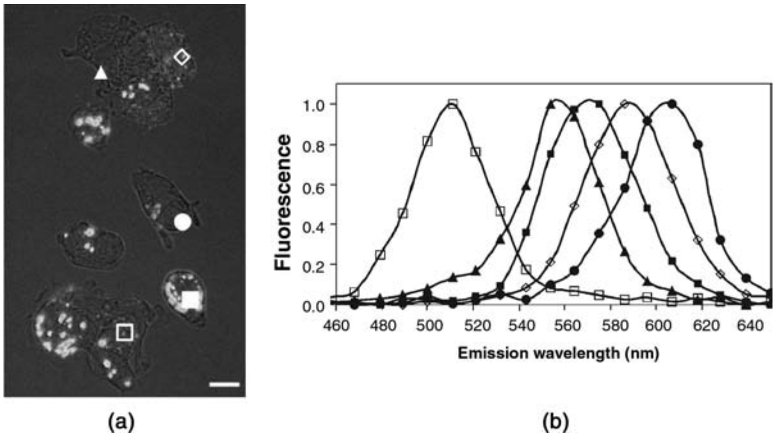


Figure 14. Spectral imaging of cells labeled by quantum dots. Cells were labeled with five different quantum dots and imaged in a multiphoton microscope. Each symbol represents a different quantum dot. The symbols on the image match the emission spectra seen on the graph. The spectral imaging set-up allows to discriminate between the different cell populations (36).

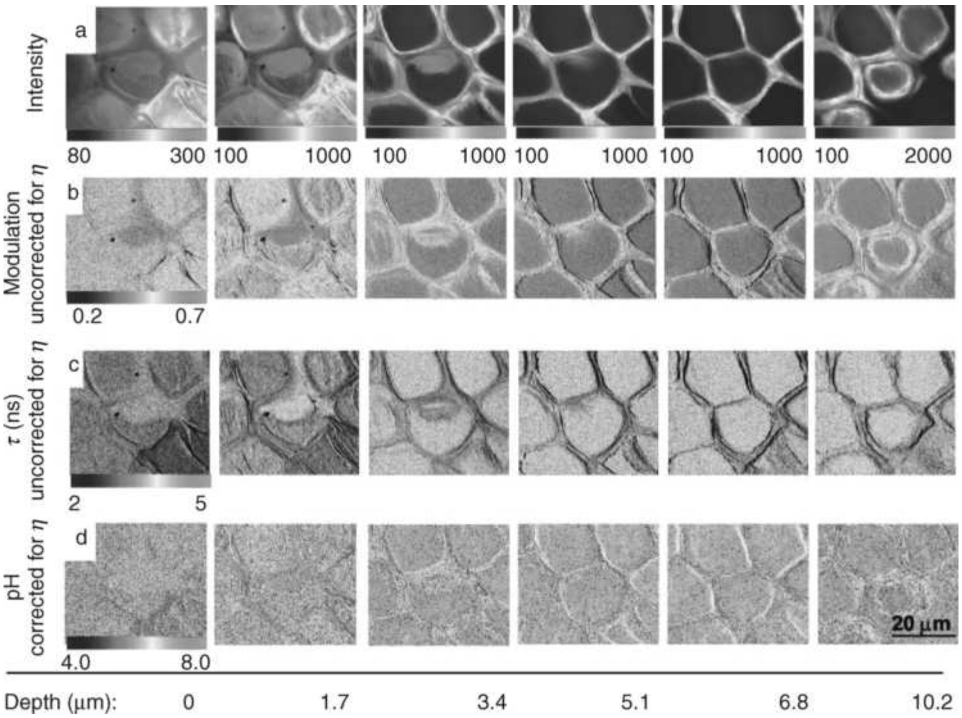


Figure 15. Quantification of the pH of the skin by lifetime imaging. (a) Intensity, (b) modulation, (c) lifetime, and (d) pH maps of a mouse skin at different depth. The lifetime measurements allow a determination of the pH independently of the intensity changes recorded between different imaging depth (37).

samples, a frequency domain approach using modulated detectors can also be used to measure the lifetime.

To measure the pH in the skin of a mouse, Clegg and co-workers (37) used a modified fluorescein probe whose lifetime varies from 2.75 ns at pH 4.5 to 3.9 ns at pH 8.5. As they image deeper in the skin, they observe that the average pH increases from 6.4 at the surface up to >7 at 15 μm depth. The extracellular space is mostly acidic (pH 6), while the intracellular space is at neutral pH. Typically, pH is measured in solution by correlating fluorescence intensities with specific pH levels. This approach is not suitable for tissues, as in the skin, since the dye is unevenly distributed throughout the tissue (Fig. 15) due to differential partitioning. A measurement of pH based on fluorescence lifetime is not dependent on probe concentration and thus the pH can be measured in the intra and extracellular space at various depths in the skin.

Polarization Microscopy

Polarization microscopy is a technique that provides information about the orientation or the rotation of fluorophores. Linearly polarized excitation light results in preferential excitation of molecules with their transition dipole aligned along the polarization. If the molecule is in a rigid environment, the emitted fluorescence will mostly retain a polarization parallel to the excitation light. However, if the molecule has time to rotate before it emits a photon, this will randomize the emission polarization. The anisotropy r is a ratio calculated from the intensity parallel $I_{\|}$ and perpendicular $I_{\perp}$ to the incident polarization and is a measure of the ability of the molecule to rotate.

$$r = \frac{I_{\|} - I_{\perp}}{I_{\|} + 2I_{\perp}} \tag{22}$$

This ratio is mostly governed by two factors, which are the fluorescence lifetime τ and the rotational correlation time θ.

$$r = \frac{r_0}{1 + (\tau/\theta)} \tag{23}$$

where r_0 is the fundamental anisotropy. Molecules with a short fluorescence lifetime and a long rotational correlation time ($\tau < \theta$) will have a high anisotropy. In the opposite case, where molecules can freely rotate during the time they reside in the excited state, the anisotropy will be low. An approximate measurement of the mobility of a molecule can be obtained by exciting the sample at different polarization angles. A proper measurement of the anisotropy requires both a linearly polarized excitation light source and the detection of the parallel and perpendicular component of the fluorescence light using a polarizer. This technique has been used to measure viscosity and membrane fluidity *In vivo*. It has been applied to quantify enzyme kinetics, relying on the fact that the cleavage of a fluorescently labeled substrate leads to a faster tumbling and thus a decrease in anisotropy.

Goldstein and co-workers (38) used polarization microscopy at the single-molecule level to study the orientation of the kinesin motor on a microtubule. A thiol reactive rhodamine dye was attached to cysteines on the motor protein. Microtubules decorated with the modified kinesin were imaged under a different polarization angle. In the presence of adenosine monophosphate (AMP)–(PNP) [a nonhydrolyable analogue of adenonine triphosphats (ATP)], the fluorescence intensity depends strongly on the angle of polarization of the excitation light (Fig. 16) proving that the kinesin maintains a fixed orientation. In the presence of adenonine5–diphosphate (ADP), however, the anisotropy is lower (no dependence on excitation polarization angle),

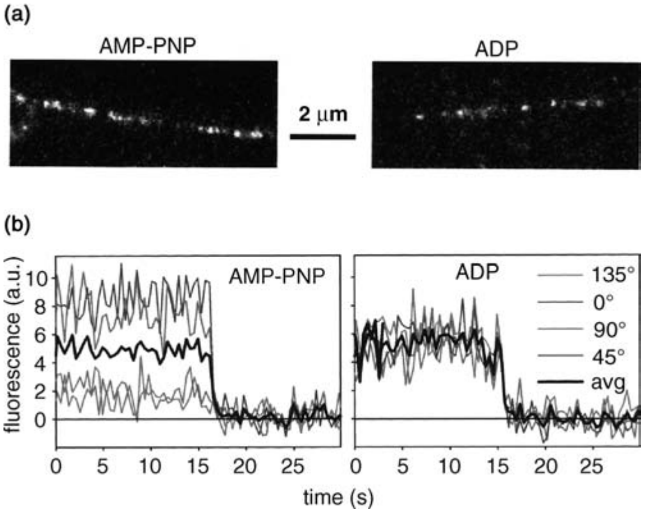


Figure 16. Mobility of single kinesin motors on microtubules probed by polarization microscopy. (a) Image of microtubules sparsely decorated with kinesin motors in presence of AMP–PNP and ADP. (b) Time course of the fluorescent intensity recorded from single molecule excited with linearly polarized light at four different angles. The large fluctuations of the fluorescence intensity as function of the excitation polarization in the AMP–PNP case demonstrate the rigidity of the kinesin motor on the microtubule (38).

leading to the conclusion that the kinesin is very mobile, while still attached to the microtubule.

Fluorescence Resonance Energy Transfer

Förster resonance energy transfer (FRET) is a technique used to monitor interaction between two fluorophores on the nanometer scale. When a dye is promoted to its excited state, it can transfer this electronic excitation by dipole–dipole interaction to a nearby molecule. Due to the nature of this interaction, Förster predicted a dependence of the FRET efficiency on the sixth power of distance that was demonstrated experimentally by Stryer with a linear polypeptide of varying length (39). The efficiency E of the process varies as function of the distance, R, between the two molecules as

$$E = \frac{R_0^6}{R_0^6 + R^6} \tag{24}$$

Where R_0 is called the Förster distance, which depends on Avogadro's number N_A, the index of refraction of the medium n, the quantum yield of the donor molecule Q_D, the orientation factor κ, and the overlap integral J.

$$R_0^6 = \frac{9000\ln(10)\kappa^2 Q_D}{128\pi^5 N_A\ n^4} J \tag{25}$$

κ represents the relative orientation of the transition dipole of the donor and acceptor molecules. In most cases, a random interaction is presumed and this factor is set to two-thirds. The overlap integral J represents the energy overlap between the emission of the donor and the absorption of the acceptor. For well-matched fluorophore pairs, R_0 is on the order of 4–7 nm.

Most FRET experiments are based on the measurement of the intensity of the donor and of the acceptor because the presence of FRET in a system is characterized by a decrease in the emission of the donor and an increase in the acceptor signal. Thus, in principle, a two color channel microscope is sufficient to follow these changes. However, experimental artifacts, such as concentration fluctuation and spectral bleed, complicate the analysis of these images and many different correction algorithms have been developed.

FRET measurements have been used in molecular studies to measure distances and observe dynamic conformational changes in proteins and ribonuclic acid (RNA). In cellular studies, FRET is often used to map protein interactions. By labeling one protein with a donor dye and its ligand with an acceptor dye, energy transfer will occur only when the two proteins are bound such that the dyes come in close proximity of each other.

The addition of fluorescence lifetime imaging provides the additional capability of retrieving the proportion of fluorophore undergoing FRET in each pixel of an image independently of concentration variations. This is possible because the fluorescence lifetime of a FRET construct is shorter than the natural decay of the dye. Thus if one has a mixture of interacting and free protein, fitting a double exponential to the fluorescence decay allows us to retrieve the proportion of interacting protein. This has been applied by Bastiaens and co-workers (40) to the study the phosphorylation of the EGF receptor ErB1. The transmembrane receptor is fused with a GFP protein. The phosphrylation of the protein is sensed by an antibody labeled with a red dye (Cy3). When the Erb1 is phosphorylated, the antibody binds to the protein and FRET occurs due to the short distance between the antibody and the GFP. The ErB1 receptors can be stimulated by EGF coated beads leading to phosphorylation and FRET. The time course of the stimulation is followed and for each cell and the fraction of phosphorylated receptors at various time interval is shown in Fig. 17. After 30 s, the FRET events are localized at discrete locations. But after 1 min, the whole periphery of the cell displays high FRET, demonstrating the lateral signaling of the receptor after activation at discrete location by EGF coated beads.

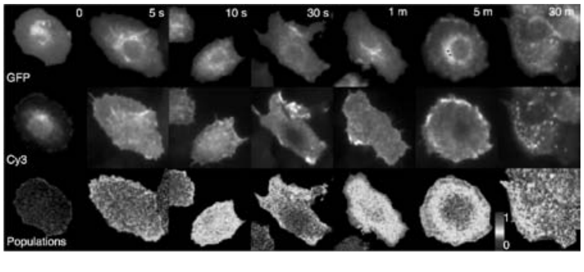


Figure 17. Time course of the phosphorylation of EGF receptors ERB1 after stimulation by EGF coated beads observed by FRET between a GFP modified EGF receptor and a phosphorylation antibody labeled with Cy3. While the GFP intensity is remains relatively constant, the concentration of the Cy3 tagged antibody clearly increases after the stimulation. This leads to an increase FRET signal as function of time (40).

CONCLUSION

The utility of fluorescence microscopy lies in its ability to study biological structure and function *In vivo*. The exquisite sensitivity and image contrast of fluorescence microscopy allow biological structures to be imaged on the submicron length scale. The greatest power of fluorescence microscopy lies in its ability to determine biochemical functions using assays based on fluorescence spectroscopy. With the availability of more versatile instruments, more fluorophores unit greater molecular and environmental specificity, the impact of fluorescence microscopy technology on biomedical science will only increase.

BIBLIOGRAPHY

1. Wang XF, Herman B. Fluorescence Imaging Spectroscopy and Microscopy. New York: Wiley; 1996.
2. Inoué S, Spring KR. Video Microscopy: the Fundamentals. New York: Plenum Press; 1997.
3. Herman B. Fluorescence Microscopy. Oxford: Bios Scientific Publishers / Springer in Association with the Royal Microscopy Society; 1998.
4. Cantor CR, Schimmel PR. Biophysical Chemistry. San Francisco: Freeman; 1980.
5. Valeur B. Molecular Fluorescence: Principles and Applications. Weinheim; New York: Wiley-VCH; 2002.
6. Lakowicz JR. NetLibrary Inc. Topics in Fluorescence Spectroscopy. Kluwer Academic; 2002.
7. Klein MV, Furtak TE. Optics. New York: Wiley; 1986.
8. Hecht E. Optics. Reading, (MA): Addison-Wesley; 2002.
9. Gu M. Advanced Optical Imaging Theory. Berlin: New York: Springer; 2000.
10. Born M,Wolf E. Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light. Cambridge: New York: Cambridge University Press; 1999.
11. Rizzuto R, Carrington W, Tuft RA. Digital imaging microscopy of living cells. Trends Cell Biol 1998;8:288–292.
12. Agard DA, Hiraoka Y, Shaw P, Sedat JW. Fluorescence microscopy in three dimensions. Methods Cell Biol 1989;30:353–377.
13. Carrington WA, et al. Superresolution three-dimensional images of fluorescence in cells with minimal light exposure. Science 1995;268:1483–1487.
14. Krishnamurthi V, Liu YH, Bhattacharyya S, Turner JN, Holmes TJ. Blind Deconvolution Of Fluorescence Micrographs By Maximum-Likelihood-Estimation. Appl Opt 1995;34:6633–6647.
15. Wilson T, Sheppard CJR. Theory and Practice of Scanning Optical Microscopy. New York: Academic; 1984.
16. Pawley JB. Handbook of Confocal Microscopy. New York: Plenum; 1995.
17. Gu M. Principles of Three-Dimensional Imaging in Confocal Microscopy. Singapore: World Scientific; 1996.
18. Masters BR. Selected Papers on Multiphoton Excitation Microscopy. Bellingham: SPIE Optical Engineering Press; 2003.
19. Amos WB, White JG. How the confocal laser scanning microscope entered biological research. Biol Cell 2003;95:335–342.
20. Denk W, Strickler JH, Webb WW. Two-photon laser scanning fluorescence microscopy. Science 1990;248:73–76.
21. Kleinfeld D, Mitra PP, Helmchen F, Denk W. Fluctuations and stimulusinduced changes in blood flow observed in individual capillaries in layers 2 through 4 of rat neocortex. Proc Natl Acad Sci U. S. A. 1998;95:15741–15746.
22. Axelrod D. Total internal reflection fluorescence. Ann Rev Biophys Bioeng 1984;13:247–268.
23. Axelrod D. Total internal reflection fluorescence microscopy in cell biology. Traffic 2001;2:764–774.
24. Axelrod D. Selective imaging of surface fluorescence with very high aperture microscope objectives. J Biomed Opt 2001;6:6–13.
25. Mason WT. Fluorescent and Luminescent Probes for Biological Activity: a Practical Guide to Technology for Quantitative Real-time Analysis. San Diego: Academic; 1999.
26. Slavic J. Fluorescence Microscopy and Fluorescent Probes. New York: Plenum Press; 1996.
27. Tsien RY. Fluorescent probes of cell signaling. Annu Rev Neurosci 1989;12:227–253.
28. Tsien RY. Fluorescent indicators of ion concentrations. Methods Cell Biol 1989;30:127–156.
29. Bruchez Jr M, Moronne M, Gin P, Weiss S, Alivisatos AP. Semiconductor nanocrystals as fluorescent biological labels. Science 1998;281:2013–2016.
30. Chan WC, et al. Luminescent quantum dots for multiplexed biological detection and imaging. Curr Opin Biotechnol 2002;13:40–46.
31. Michalet X, et al. Quantum dots for live cells, in vivo imaging, and diagnostics. Science 2005;307:538–544.
32. Zhang J, Campbell RE, Ting AY, Tsien RY. Creating new fluorescent probes for cell biology. Nat Rev Mol Cell Biol 2002;3:906–918.
33. Chalfie M, et al. Green Fluorescent Protein as a Marker for Gene Expression. Science 1994;263:802–805.
34. Lakadamyali M, Rust MJ, Babcock HP, Zhuang X. Visualizing infection of individual influenza viruses. Proc Natl Acad Sci U. S. A. 2003;100:9280–9285.
35. Gaietta G, et al. Multicolor and electron microscopic imaging of connexin trafficking. Science 2002;296:503–507.
36. Voura EB, Jaiswal JK, Mattoussi H, Simon SM. Tracking metastatic tumor cell extravasation with quantum dot nanocrystals and fluorescence emissionscanning microscopy. Nat Med 2004;10:993–998.
37. Hanson KM, et al. Two-photon fluorescence lifetime imaging of the skin stratum corneum pH gradient. Biophys J 2002;83:1682–1690.
38. Sosa H, Peterman EJ, Moerner WE, Goldstein LS. ADP-induced rocking of the kinesin motor domain revealed by single-molecule fluorescence polarization microscopy. Nat Struct Biol 2001;8:540–544.
39. Stryer L, Haugland RP. Energy Transfer: a spectroscopic ruler. Proc Natl Acad Sci U. S. A. 1967;58:712–726.
40. Verveer PJ, Wouters FS, Reynolds AR, Bastiaens PIH. Quantitative imaging of lateral ErbB1 receptor signal propagation in the plasma membrane. Science 2000;290:1567–1570.

See also FLUORESCENCE MEASUREMENTS; MICROSCOPY, CONFOCAL

MICROSCOPY, SCANNING FORCE

EWA P. WOJCIKIEWICZ
KWANJ JOO KWAK
VINCENT T. MOY
University of Miami
Miami, Florida

INTRODUCTION

Recent advances in technology have allowed us to study our world at molecular, even subatomic, resolution. One of the devices in the forefront of such studies is the atomic force microscope (AFM), which is a relatively complex device with two major applications. It can be used as an imaging device, which allows for the acquisition of atomic-level images of biological structures as well as to measure forces of interactions between two opposing surfaces down to the single-molecule level.

Imaging AFM

The AFM was originally designed as an imaging tool (1). It was modified from the design of the scanning tunneling microscope (STM). The AFM acquires topographic images by methodically scanning a sample with a flexible probe, called a cantilever, which bends according to the contours of the sample's surface. The bending of the cantilever is translated into an image map, which reveals the height differences in the surface being scanned. It is possible to image biological samples under physiological conditions as imaging can be done in both air and liquid. The resulting resolution of such maps is at the atomic level (2,3).

The imaging AFM has been used to image many biological samples ranging from genetic material to cells to bone. A few of these studies will be highlighted. One of the earliest biological materials to be imaged was DNA, which has been imaged in many forms to date, including double- and single-stranded forms as well as more complex structures. The AFM has also been used for many applications including DNA sizing, previously only achieved using gel electrophoresis, DNA mapping, hybridization studies, and examinations of protein-DNA interactions (4). AFM studies of RNA were also conducted. Unlike DNA, which mainly forms a double-helical structure, RNA has the ability to form more advanced structures that do not rely solely on Watson–Crick base pairing. One example are the so-called kissing-loop structures imaged by Hansma et al. (4) (Fig. 1). Not only was the AFM used in imaging of such structures, many of them 3D, but also played an important role in designing them (5). Unlike other imaging techniques, AFM sudies can be done under physiological conditions allowing for the imaging of biological processes. Images of transcription complexes have been obtained, for example, *E.coli* RNA polymerase in complex with DNA. These studies are the only of their kind that can answer certain specific questions as to how the RNA transcription process takes place. One is able to visualize how the DNA does not get entangled in the nascent RNA strands. Such studies detailing the structure-function relationship of the transcription process are key in furthering our understanding of gene expression (6,7).

Also, imaging of cells was conducted to examine the structure of the cellular cortex in detail. The cell cytoskeleton is known to be involved in affecting cell shape as well as movement and other cellular responses to biochemical and biophysical signals. At present, relatively little is known about the mechanical organization of cells at a subcellular level. Pesen et al. (8) studied the cell cortex of bovine pulmonary artery endothelial cells (BPAECs) using AFM and confocal fluorescence microscopy (CFM). They were able to identify a coarse and fine mesh that make up the cortical cytoskeleton. These two types of mesh appear to be intertwined (Fig. 2) (8). Such details are not distinguished in imaging studies using fixed cells.

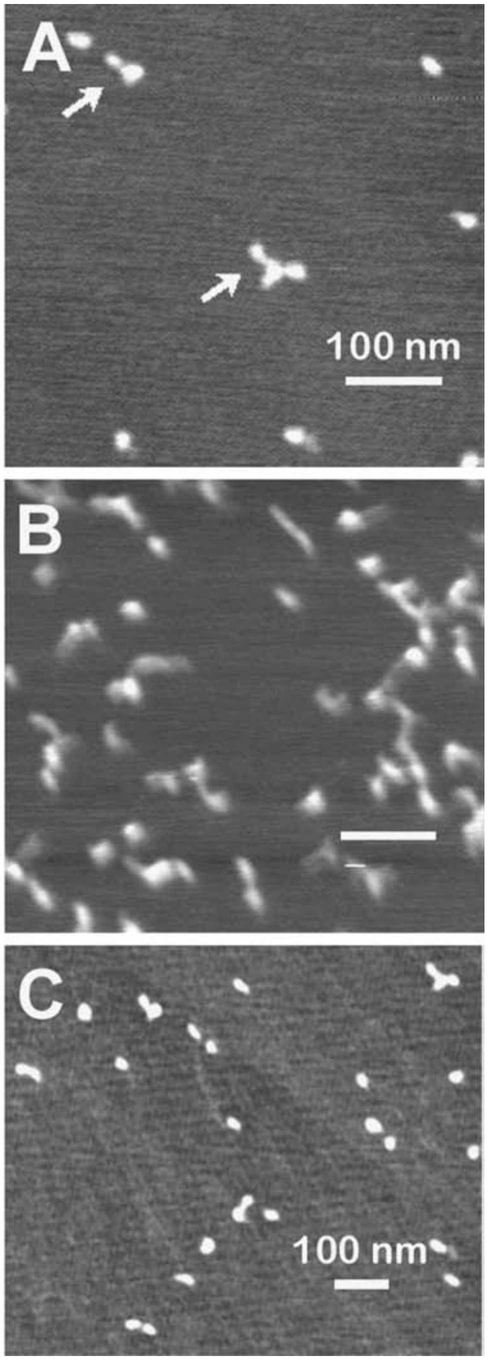


Figure 1. Kissing-loop RNA structures. (a,c) Kissing-loop RNAs at low concentrations. The three arms of the individual RNAs are visible. (B) Kissing-loop RNAs at a concentration 10-fold higher than in a and c. The individual structures are less well-defined. Scale bars = 100 nm for all images.

Other imaging studies have looked at tendon and bone, both of which are composed of type I tropo-collagen, which was done by acquisition of high resolution AFM images of type I collagen in conjunction with force spectroscopy studies, namely protein unfolding, which is described in the following section. Figure 3 reveals these high resolution collagen type I images. They were acquired using two different concentrations of collagen, which resulted in

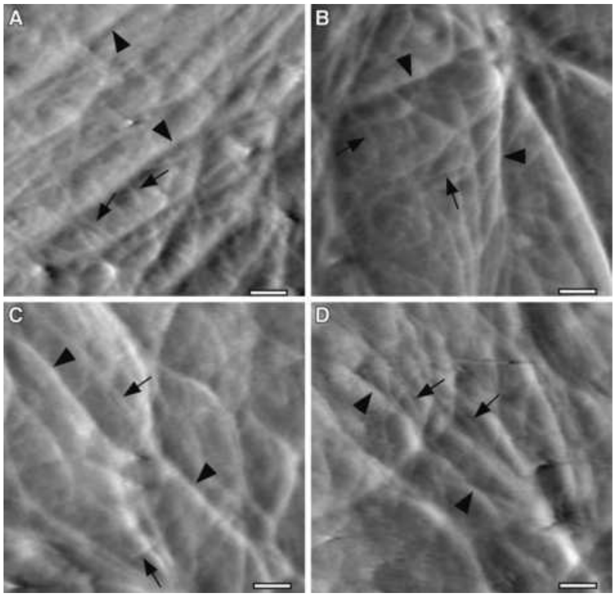


Figure 2. AFM Images of the cortical mesh of bovine pulmonary artery endothelial cells (BPAEC). (a–d) High magnification deflection AFM images of the cortical mesh of living BPAECs in a physiological saline. The filamentous mesh appears to be organized on two length scales, with coarse mesh (arrowheads) and fine mesh filaments (arrows). The two meshes are likely to be intertwined, although it is possible that the fine mesh is layered over the coarse mesh. Lateral resolution in these images is ~125 nm.

slightly different orientations of collagen: random at the lower concentration (Fig. 3a) and oriented unidirectionally in the higher concentration (Fig. 3b). In these studies, the AFM was used to investigate the mechanical properties of this collagenous tissue, which are altered in diseases such as osteoporosis. Being familiar with such properties is important for gaining further understanding as well as preventing and curing bone diseases (9).

Force Spectroscopy

The AFM can also be operated in the force scan mode, which allows for the measurement of adhesion forces between receptors and their corresponding binding partners, or ligands. In studies of ligand-receptor forces, the receptor is immobilized on the surface of a flexible AFM cantilever whereas the ligand is attached to a suitable substrate. The deflection of the cantilever during the approach and withdrawal of the cantilever from the substrate allow for the force of the interaction to be measured. These type of experiments provide information simulating the influence of internal and external forces that these receptors would experience in the body, for example, the shear stress experienced by a blood cell attached to the endothelium while blood rushes past it. Such information was previously unavailable when receptor-ligand interactions were examined using more traditional biochemical techniques. AFM has made it possible to acquire measurements that reveal the mechanical properties of biomolecules under applied force. Measurements of the unbinding force of a single ligand-receptor interaction can now be acquired with the AFM (10–12).

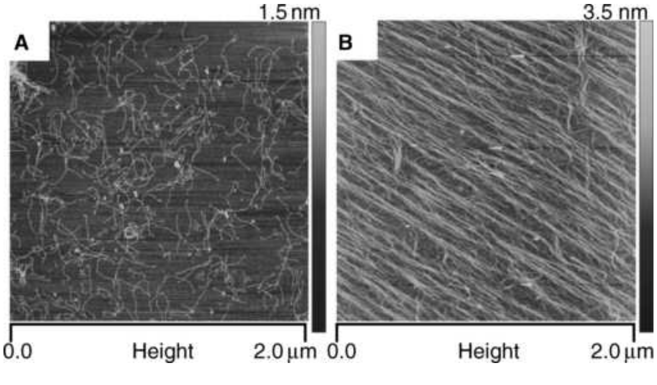


Figure 3. Topographical images (height; tapping mode in air) of type I collagen monomers on a mica substrate. (a) low (1 μg/ml) concentration of collagen, (b) high (10 μg/ml) concentration of collagen.

The AFM can also be used in adhesion studies involving whole cells (13,14). In these studies, the interaction between a cell expressing a particular receptor of interest and its ligand protein or another cell expressing the ligand is measured. The cell adhesion experiments allow for the acquisition of both single-molecule measurements, like in the above-mentioned studies, as well as multiple-bond interactions. The advantage of using the AFM in cell adhesion studies is the high specificity and wealth of information that is obtained. The AFM force scans provide information about the individual bond strengths as well as the force and work that is required to separate the entire complex. Combining single-molecule and multiple-bond data allows us to describe the thermodynamic model of the separation of a particular complex in addition to the mechanism of its action on the cellular scale (15,16).

The AFM can also serve as a microindenter that probes soft samples, including cells revealing information about their mechanical properties. The mechanical properties of cells play an important role in such essential physiological processes such as cell migration and cell division. Understanding these properties can later help scientists to identify when certain processes may be taking place. The mechanical properties of cells are chiefly determined by their actin cytoskleleton, which is the cell's "backbone." This type of information, which cannot be obtained using standard cell biology methods, allows for the estimation of the Young's modulus of living cells (16). The Young's modulus is a calculated value, which provides information regarding the compliance or elasticity of a cell. Such experiments may be done with either the imaging AFM or using force spectroscopy. Manfred Radmacher's group has conducted such measurements with an imaging AFM in force mapping mode. The advantage of such measurements is the wealth of information that they provide. These experiments reveal not only the elastic properties of the cells being examined but also topographical information. They can also be performed in conjunction with video microscopy to further confirm what one is visualizing and that cells are not undergoing damage (17,18). In force spectroscopy experiments, the Young's modulus is obtained by poking the cell cantilever tip. This type of information can be correlated with adhesion data to determine whether elasticity changes in response to drugs or other stimulation have an effect on the strength of cell adhesion (19).

Another application allows scientists to study protein folding. Proteins are composed of amino acid chains, which constitute the primary protein structure. These amino acid chains have to undergo folding into tertiary and secondary, 3D structures, which is essential for the proper functioning of these proteins. Recently, advances in AFM have made it possible to study this fascinating biological process and bring new insight into the energy landscapes of protein folding. Proteins involved in mechanical functions are composed of multiple domains, which fold individually. One of these domains is fibronectin. The most common form of fibronectin is fibronectin type III (FN-III), which is found in an estimated 2% of all animal proteins and has thus been studied extensively. FN-III domains are found in the muscle protein titin, which is responsible for the passive elasticity of muscle. These domains have been found to unravel during forced extension. Understanding the forces required in unfolding events as well as the time it takes for unfolding to happen can be critical for the physiological functions of mechanical proteins such as titin (9,20–22).

Now that a brief overview of the potential applications of the AFM has been provided, it is important to understand the principles behind its operation. The following section focuses on the theory of data acquisition using AFM as well as descriptions of the equipment itself. The last section of this article provides a more in-depth evaluation of the technique in addition to discussing the most recent advances in the field.

THEORY AND EXPERIMENTAL APPROACH

This section focuses on the theory and experimental approaches of AFM. The section begins with a description of the imaging AFM as well as its different modes of operation, which allow for its applications in various experimental protocols. Later, force spectroscopy is described. Focus is placed on the force apparatus used in our laboratory, which relies on the same basic principals as the commercially available AFMs. In addition to a description of its operation, this section also discusses the different applications of the force apparatus.

Imaging AFM

Optical Beam Deflection. An AFM has a force sensor called a cantilever to measure the forces between a sharp tip and the surface (23). Unlike the optical microscope that relies on 2D images, the images acquired with the AFM are obtained in three dimensions: the horizontal *xy*-plane and the vertical *z*-direction. As shown in Fig. 4, the tip at the end of the cantilever is brought in close proximity to the sample mounted on a piezoelectric element. The AFM can be compared with a record player such as an old stylus-based instrument (1). It combines the principles of a scanning tunneling microscope (STM) and the stylus profiler. However, the probe forces in the AFM are much smaller than those ($\sim 10^4$ N) achieved with a stylus profiler.

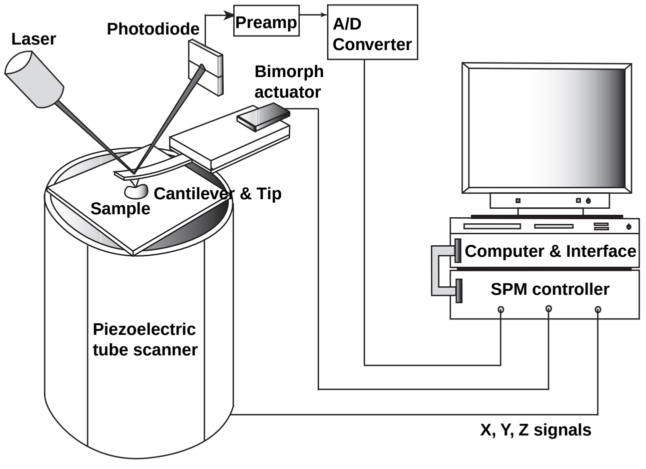


Figure 4. Schematic illustration of an AFM. A force sensor consists of a flexible cantilever with an extremely sharp tip at its end. A ceramic (Si_3N_4) or semiconductor (Si) tip on the cantilever can be brought into close proximity to a sample surface. As the tip is close to the sample surface, it either continuously touches or periodically vibrates on the surface, and bends or changes in its vibration amplitude and frequency. A laser spot is reflected off the top of the cantilever. When the cantilever bends, the laser light is deflected onto a two-panel photodiode. The detected signals are amplified and transformed by electronic circuits and sent to an SPM controller. The SPM controller, the computer, and their interfaces generate an image.

The AFM belongs to the family of scanning probe microscopes (SPMs). Like all other SPMs, the AFM uses an extremely sharp tip that moves over the sample surface with a raster scan (like the movement of the electron beam on the TV screen). In the first AFM, the bending of the cantilever was detected using an STM, but now a sensitive and simple optical method is used in most AFMs (24). As shown in Fig. 4, a laser beam is reflected off the cantilever onto a two-panel photodiode. As the cantilever bends, the position of the reflected laser light changes. Measurements are obtained as a result of the differences in the signal between the two segments of this photo-detector.

Feedback Operation. The AFM uses a piezoelectric element to position and scan the sample with high resolution. A tube-shaped piezoelectric ceramic that has a high stability is used in most SPMs. Application of voltage results in the stretching or bending of the piezoelectric tube, allowing it to move in all three dimensions and to position the cantilever probe with very high precision. For example, by applying a voltage to one of the two electrodes (*xy*-axis) the tube scanner expands and tilts away from a center position (*xy*-origin). A corresponding negative voltage applied to the same electrode causes the tube scanner contract, resulting in movements on the *xy*-plane relative to the origin. The magnitude of the movement depends on the type of piezoelectric ceramic, the shape of the element, and the applied voltage.

Feedback control is used for many common applications, such as thermostats, which are used to maintain a particular temperature in buildings, and autopilot, commonly used in airplanes. In the AFM, a feedback loop is used to keep the force acting on the tip in a fixed relationship with the surface while a scan is performed. The feedback loop consists of the piezoelectric tube scanner, the cantilever and tip, the sample, and the feedback circuit. The feedback circuit consists of proportional and integral gain controls and provides an immediate response to scanning parameter changes. A computer program acts as a compensation network that monitors the cantilever deflection and attempts to keep it at a constant level.

Contact Mode (Static Mode). The AFM operates by measuring the intermolecular forces between the tip and sample. The most common method used in imaging AFM is contact mode, where the piezoelectric element slightly touches the tip to the sample. The experimental setup is shown in Fig. 4. As a result of the close contact, the tip and sample remain in the repulsive regime of the tip-sample interaction shown in Fig. 5. Thus, the AFM measures repulsive force between the tip and sample. As the raster scan moves the tip along the sample, the two-panel photodiode measures the vertical deflection of the cantilever, which reveals the local sample height. Each contour of the surface results in a movement of the tip in the *xyz*-direction, resulting in a change in the deflection angle of the laser beam. This change is measured through the photodiode and translated finally to an image.

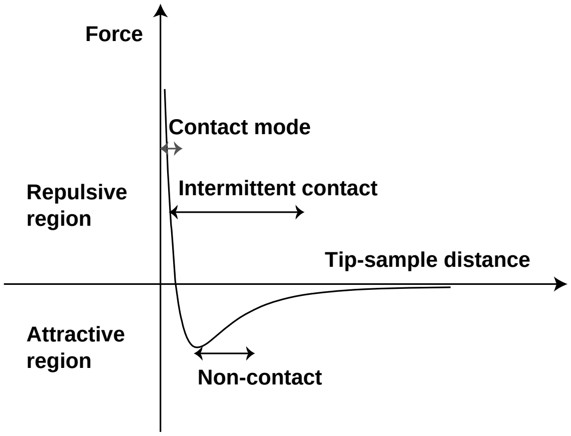


Figure 5. Relationship between the operating modes of AFM and the force regions. The x-axis of the graph is the tip-sample distance and the y-axis is the force or potential. The graph shows the force or potential, as a function of the distance, simply calculated by the Lennard–Jones potential and the DMT approximation. In contact mode AFM, the tip-sample interaction lies in the repulsive region of the curve above the x-axis. As the cantilever is pushed upward, a resulting restoring force occurs, which can be given by Hooke's Law. The difference between this graph and the measured force-distance curve occurs when the force measurement is not static but dynamic and is quickly affected by the capillary force in the thin surface water layer. In intermittent contact mode, the cantilever is operated by a vibration of an amplitude of 10 to 100 nm. The tip is still in contact with the surface, but it just contacts or "taps" on the surface for a very small fraction of its oscillation period. In this operation mode, the tip-sample interaction is broad, ranging from the repulsive region of the curve to the attractive region due to the long-range van der Waals force. The tip-sample interaction in noncontact mode is much weaker than the one in contact and intermittent contact mode. The force between the tip and sample is several orders of magnitude weaker than in the contact regime. As the cantilever in noncontact mode is vibrated near its resonance frequency with an amplitude less than 10 nm, the spacing between the tip and sample is on the order of several to tens of nanometers.

Tip-sample Interaction: The cantilever in the AFM is a critical component (1). The force produced by a spring always tends to restore the spring to its equilibrium position. When the spring is pushed upward by a distance z, it has to be pulled downward. This restoring force is given by Hooke's Law as:

$$F(z) = -k * (z - z_0) \qquad ()$$

where k is a spring constant and depends on the material and dimensions of the cantilever, z is the vertical position of the cantilever, and z_0 is the equilibrium position. As a cantilever with a spring constant of 0.1 newton/meter (N/m) is moved by 1 nm, the resulting force is 0.1 nanonewton (nN). The first tip used by the inventors of the AFM was made from diamond glued to a lever of gold foil (1). Microfabricated cantilever tips are now commercially used.

Electromagnetic forces determine the properties of solids, liquids, and gases; the behavior of particles in solution; and the organization of biological structures (25). These forces are also the source of all intermolecular interactions including covalent bonds, Coulomb forces, ionic forces, ion-dipole interaction, and dipole-dipole interaction. In the AFM, the intermolecular interactions between the tip and the sample surface include van der Waals forces, electrostatic forces, water capillary force, and material properties including elasticity. The most common force in the tip-sample interaction is the van der Waals force. The force is calculated using the Lennard–Jones potential, which combines the attractive van der Waals and the repulsive atomic potentials (25). The force depends on the distance between the tip and the sample, as shown in Fig. 5. This calculated force is an estimate of the van der Waals forces and is usually a few nanonewtons in magnitude.

Force-distance Curve: Since the invention of the AFM, many researchers have used it to measure the tip-sample interaction force on the atomic scale. The AFM records the force as the tip is brought in close proximity to the sample surface, even indented into the surface, and then pulled off. The measured force curve is a plot of cantilever deflection versus the extension of the z-piezoelectric scanner (z-piezo). A force-distance curve is a kind of interpretation of the force measurements. It needs a simple relationship between the cantilever deflection and the tip-sample distance. Thus, the force-distance curve describes the tip-sample interaction force as a function of the tip-sample distance rather than as a function of the z-piezo position. It is difficult to measure the quantitative forces with this technique because the spring constant of the cantilever and the shape of the tip are not accurately known. However, this technique has been used to study adhesion, elasticity, bond rupture length, and even thickness of adsorbed layers. These studies of the fundamental interactions between the sample surfaces have extended across basic science, chemistry, biology, and even material science. The interaction force between tip and sample is typically on the order of tens of pN for biomolecular interactions. Force measurements in solution have the advantages of the AFM due to the lower tip-sample interaction.

Constant Force and Constant Height: In contact mode, the tip is scanned across the surface in contact either at a constant force or at a constant height above the sample. Constant force mode is achieved by use of a *z*-feedback loop from the deflection signal. The feedback circuits serve to maintain a constant force between the tip and the sample while the tip follows the contours of the surface. The piezoelectric tube can respond to any changes in the cantilever deflection. A computer program acts to keep the cantilever deflection at a constant level. Then, the tip-sample interaction can be kept at a predetermined restoring force. This technique is used to observe the precise topography of the sample surface. If the *z*-feedback loop is switched off, then the *z*-direction of the piezoelectric tube is kept constant, and an image is generated based on the cantilever deflection. Using constant height can be useful for imaging very flat samples.

Lateral Force Microscopy: Lateral force microscopy (LFM) is an extension of contact mode, where an additional detected parameter is the torsion of the cantilever, which changes according to the friction force (26). This lateral force induces a torsion of the cantilever, which, in turn, causes the reflected laser beam to undergo a change in a perpendicular direction to that resulting from the surface corrugation. The LFM uses a photodiode with four segments to measure the torsion of the cantilever. When the cantilever is scanned across the surface in contact, differences in friction between tip and sample cause the tip to stick-slip on the surface. This stick-slip behavior creates a characteristic saw-tooth waveform of atomic level in the friction image (27). The LFM can provide material-sensitive contrast because different components of a composite material exert different friction forces. Researchers often call this operation mode friction force microscopy (27,28). Increasing wear with decreasing sliding velocity on the nanometer scale has been observed with this technique. It has been demonstrated with LFM that, on the atomic scale, frictional properties are sensitive to changes in surface properties on chemical modification. The LFM can also be applied to chemical force microscopy (CFM) by a modified tip with chemical functionality (29). It has been demonstrated with CFM that mapping the spatial arrangement of chemical functional groups and their interactions is of significant importance to problems ranging from lubrication and adhesion to recognition in biological systems.

Capillary Force: The thin surface water layer that exists on the sample surface will form a small capillary bridge between the tip and the sample. The capillary force is important when the AFM is operated in air. Examine the effect of surface tension on AFM measurements. At the moment of tip contact with a liquid film on a flat surface, the film surface reshapes producing a ring around the tip. The water layer wets the tip surface because the water-tip contact (if it is hydrophilic) is energetically advantageous as compared with the water-air contact. If the tip radius is 10 nm and the contact angle is small (i.e., hydrophilic), a capillary force of about 10 nN can result. Thus, the capillary force is the same order of magnitude as the van der Waals interaction. An AFM tip has been used to write alkanethiols with a 30 nm line-width resolution on a gold thin film in a manner analogous to that of a dip pen (30). Recently, this dip-pen nanolithography has also been applied to direct nanoscale patterning of biological materials such as DNA, peptides, and proteins on glass substrates.

Vibration Mode (Dynamic Mode). In dynamic mode, the cantilever is oscillated close to its resonance frequency. This vibration mode operates at a frequency-modulation (FM) mode or the more common amplitude-modulation (AM) mode, which are basically the same as the frequencies used in radio communication. In the FM mode, a *z*-feedback loop keeps a constant force between the tip and the sample while the tip follows the contours of the surface by maintaining the resonance frequency. In the AM mode, the *z*-feedback loop keeps the constant tip-sample interaction by maintaining the amplitude of oscillation.

Intermittent Contact Mode: The cantilever in dynamic mode can easily be vibrated by a piezoelectric ceramic called a bimorph actuator. In air, the cantilever is oscillated close to its resonance frequency and positioned above a sample surface. When the vibrating cantilever comes close to the surface, its oscillation amplitude may change and can be used as the control signal. In this AM mode, the tip is still in contact with the surface, but it just contacts or "taps" on the surface for a very small fraction of its oscillation period. This operation mode is best known as tapping mode in commercial AFMs and, more generally, as intermittent contact mode.

As a raster scan moves the tip on the sample, the four-segment photodiode measures the vibration signal of the cantilever. The detected signal can be changed to root mean-square values by an analog-to-digital converter. In constant force mode, the *z*-feedback loop adjusts so that the averaged amplitude of the cantilever remains nearly constant. Each contour of the surface causes a movement of the tip in the *xyz*-direction, resulting in a change of the oscillation amplitude of the cantilever. This change is measured through a photodiode and finally translated to an image. In air, friction forces due to the surface water layer are dramatically reduced as the tip scans over the surface. Tapping mode may be a far better choice than contact mode for imaging of biological structures due to their inherent softness. In tapping mode, the cantilever can be vibrated at an amplitude of less than 100 nm. Additionally, changes in the phase of oscillation under tapping mode can be used to discriminate between different types of materials on the surface.

Tip-Sample Interaction: The mechanical resonance of the cantilever plays a major role in the response of the system for an interaction between a tip mounted on a vibrating cantilever and a non-homogeneous external force (23). Although basic equations governing the operation of a bimorph actuator used to vibrate the cantilever are not introduced here, the position of the bimorph is given by:

$$u = u_0 + A_{ex}\cos(\omega t+)$$

where u_0 is the equilibrium position and the excitation is done with amplitude A_{ex}, a frequency ω, and a phase shift ϕ. The fundamental resonance frequency of the cantilever can be approximately calculated from equating its strain energy at the maximum deflection to the kinetic energy at the point of zero deformation. A more accurate method, which takes into consideration all the resonance frequencies of the cantilever together with the modes of vibration, can be obtained by solving the equation of motion subject to the boundary conditions (23). A basic equation to describe the motion of the cantilever is briefly introduced. If the tip-sample interaction is uniform and includes dissipative force in Newton's second law, the vibration system including the cantilever can be described as follows:

$$F(z) = k(z-u) + y(dz/dt) + m(d^2z/dt^2) \qquad ()$$

where *F(z)* is the tip-sample interaction force, k is a spring constant of the cantilever, z is the vertical position of the cantilever, u is the motion of the bimorph, γ is the dissipation term (i.e., the friction coefficient of the material or the environment), and m is the effective mass of the cantilever. For the constant amplitude mode, we assume that the frictional force γ (dz/dt) is compensated for by the driving force $F_{ex} = k\,A_{ex}\,\cos(\omega t + \phi)$. Then, the equation of motion is reduced to $F(z) = k\,z + m(d^2z/dt^2)$. If a strong tip-sample interaction occurs only at the point of contact, the motion of the cantilever tip can be almost perfect harmonic oscillation, $z = zo + A\sin\omega t$.

Resolution and Tip Effects: The resolution obtained by an AFM depends greatly on the sharpness of the cantilever tip. Broadening effects usually develop when imaging biological structures having extremely small surface features like a DNA strand (4). If a tip with a radius of curvature of about 20 nm is used to image DNA on a substrate surface, the observed width is about 20 nm, which is considerably greater than the expected value of 2.4 nm deduced from the van der Waals radii of DNA. When the tip radius is comparable with the size of the feature being imaged, it is important to evaluate the radius of the tip end. As such, the development of sharper tips is currently a major concern for commercial vendors, which is also of interest for biologists whose work would greatly benefit from much faster scanning. Recently, improvement of the scanning speed in AFM is one of the most important topics. The tip-sample interaction also tends to distort biological structures because they are relatively soft (31).

Phase Imaging: Phase imaging is an extension of tapping mode based on the measurement of the cantilever phase lag (32). The dependence of phase angles in tapping mode AFM on the magnitude of tip-sample interactions has been demonstrated. The phase images of several hard and soft samples have been recorded as a function of the free amplitude and the reference of the tapping amplitude. It is thought that the elastic deformation associated with the tip-sample repulsive force can be estimated by the repulsive contact interaction. In many cases, phase imaging complements the LFM and force modulation techniques, often providing additional information along with a topographic image. Phase imaging like LFM can also be applied to CFM by using a modified tip with chemical functionality.

Pulsed Force Mode: Pulsed force mode (PFM) is a nonresonant and intermittent contact mode used in AFM imaging (33). It is similar to tapping mode in that the lateral shear forces between the tip and the sample are also reduced. In contrast to tapping mode, the maximum force applied to the sample surface can be controlled, and it is possible to measure more defined surface properties together with topography. This mode is similar to the force modulation techniques of CFM in that a chemically modified tip is used. A series of pseudo force-distance curves can be achieved at a normal scanning speed and with much lower expenditure in data storage. A differential signal can be amplified for imaging of charged surfaces in terms of an electrical double-layer force.

Noncontact Mode: A reconstructed silicon surface has been imaged in a noncontact mode by AFM with true atomic resolution (34). The operation of the AFM is based on bringing the tip in close proximity to the surface and scanning while controlling the tip-sample distance for the constant interaction force. The tip-sample interaction forces in noncontact mode are much weaker than those in contact mode, as shown in Fig. 5. The cantilever must be oscillated above the sample surface at such a distance as is included in the attractive regime of the intermolecular force. Most surfaces are covered with a layer of water, hydrocarbons, or other contaminants when exposed to air, which makes it very difficult to operate in ambient conditions with noncontact mode. Under ultrahigh vacuum, clean surfaces tend to stick together, especially when the materials are identical. The FM mode used in noncontact mode can keep the constant tip-sample interaction by maintaining the resonance frequency of oscillation through the *z*-feedback loop. Nearly ten years following the invention of the AFM, a few groups achieved true atomic resolution with a noncontact mode (35). After that, several groups succeeded in obtaining true atomiclevel resolution with noncontact mode on various surfaces. Many important yet unresolved problems, such as determining the tip-sample distance where atomic-level resolution can be achieved, still remain. Experimentally, atomic-level resolution can be achieved only between 0 and 0.4 nm. A stiff cantilever vibrates near resonance frequency (300–400 kHz) with amplitude of less than 10 nm.

In covalently bound materials, the charge distribution of surface atoms reflects their bonding to neighboring atoms (36). These charge distributions have been imaged by noncontact mode with a light-atom probe such as a graphite atom. This process revealed features with a lateral distance of only 77 picometers (pm). However, all of the atomic-scale images have been generated in ultrahigh vacuum, which has limited applications in biology. Recently, several groups have reported obtaining atomic-scale images with FM mode in ambient conditions and liquid environments. In the near future, true atomic-level imaging by AFM will be commercially available in various environments.

Force Spectroscopy

Equipment

AFM Instrumentation. The AFM that is used in the author's laboratory is a homemade modification of the standard AFM design that is used for imaging and is shown in Fig. 5 (19). It operates on the same basic principles as a commercial AFM. In the author's design, improvement of the signal quality by reducing mechanical and electrical noise and improvement of the instrument's sensitivity by uncoupling the mechanisms for lateral and vertical scans was achieved. The cantilever is moved vertically up and down using a piezoelectric translator (Physik Instrumente, model P-821.10) that expands or contracts in response to applied voltage. The vertical range of the piezo is 0–15 μm. A dish coated with substrate is placed below the cantilever, and the cantilever with a cell or protein attached can be lowered onto that dish using the piezo allowing for the receptor-ligand interaction to take place. During the acquisition of a force scan, the cantilever is bent (Fig. 6) causing the beam of a 3 mW diode laser (Oz Optics; em. 680 nm) that is focused on top of the cantilever to be deflected. A two-segment photodiode (UDT Sensors; model SPOT-2D) monitors these deflections of the laser beam. An 18 bit optically isolated analog-to-digital converter (Instrutech Corp., Port Washington, NY) then digitizes the signal from the photodiode. Custom software is used to control the piezoelectric translator and to time the measurements. The AFM is shielded inside of an acoustic/vibration isolation chamber in order to reduce vibration and aid in keeping a stable temperature. The detection limit of the AFM system is in the range of 20 pN.

Cantilever Calibration. It is necessary to determine the spring constant of the cantilever k_C (i.e., $F = k_C x$) in order to translate the deflection of the cantilever x to units of force F. Calibrating the cantilever can be achieved through theoretical techniques that provide an approximation of k_C (37) or through empirical methods. Using empirical methods to determine k_C involves taking measurements of cantilever deflection with application of a constant known force (38) or measuring the cantilever's resonant frequency (39). The method the author's use for calibrating cantilevers is based on Hutter and Bechhoefer (39). The author's use triangleshaped unsharpened gold-coated silicon-nitride cantilever tips that have spring constants ranging from 10 mN/m to 50 mN/m for ligand-receptor force measurements (TM Microscopes, Sunnyvale, CA). The cantilever tip can be treated as a simple harmonic oscillator whose power spectrum of thermal fluctuation can be used to derive the spring constant, which can be achieved by raising the cantilever a few microns from the surface of the experimental dish and monitoring its natural vibrational frequency for 2–3 s. Each vibration mode of the cantilever receives the thermal energy commensurated to one degree of freedom, $k_B\,T/2$. The measured variance of the deflection $\langle x\rangle^2$, can then be used to calculate the spring constant (i.e., $k_B T = k_C \langle x\rangle^2$, where k_B and T are Boltzmann's constant and temperature, respectively). To separate deflections belonging to the basic (and predominant) mode of vibration from other deflections or noise in the recording system, the power spectral density of the temperature-induced deflection is determined. The spring constant is estimated using only the spectral component corresponding to the basal mode of vibration. The spring constant can be calibrated in either air or solution using this approach. The calculated spring constant k_C can then be used to calculate rupture force F by $F = k_C CD\,V$. DV is the change in voltage detected by the photodiode just prior to and immediately after the rupture event. C is a calibration constant that relates deflection and photodiode voltage and is determined from the deflection of the cantilever when it is pressed against a rigid surface, such as the bottom of a plastic petri dish (19).

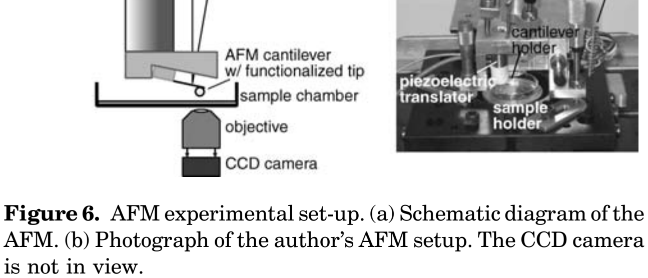


Figure 6. AFM experimental set-up. (a) Schematic diagram of the AFM. (b) Photograph of the author's AFM setup. The CCD camera is not in view.

Applications

Receptor-Ligand Adhesion Measurements. *Bell Model:* AFM force measurements (Fig. 7) of ligand-receptor interactions can be used to determine the dynamic strength of a complex and characterize the changes in free energy that the particular complex undergoes (i.e., energy landscape) during its breakage. The Bell model can be used to interpret these measurements (40). The Bell model is based on the assumption that the application of an external mechanical force to a receptor-ligand interaction bond will reduce the activation energy that needs to be overcome in order to break this bond. This unbinding force should increase with the logarithm of the rate at which an external mechanical force is applied toward the unbinding of adhesion complexes (i.e., loading rate), which was confirmed by a number of studies. For example, studies using the biomembrane force probe (BFP) (40) and the AFM have shown that increases in loading rate cause an increase in rupture force between individual complexes of streptavidin/biotin (12,15,41).

AFM Measurements of Adhesive Forces: In order to carry out force measurements, a cell is first attached to a cantilever tip and another cell or substrate proteins are plated on a dish. The method employed to attach cells to the cantilever tip works best on nonadherent items. A cell is attached to the AFM cantilever via concanavalin A (con A)-mediated linkages (15). Most cells have receptors for con A on their surface and will attach to the tip. To prepare the con A-functionalized cantilever, the cantilevers are soaked

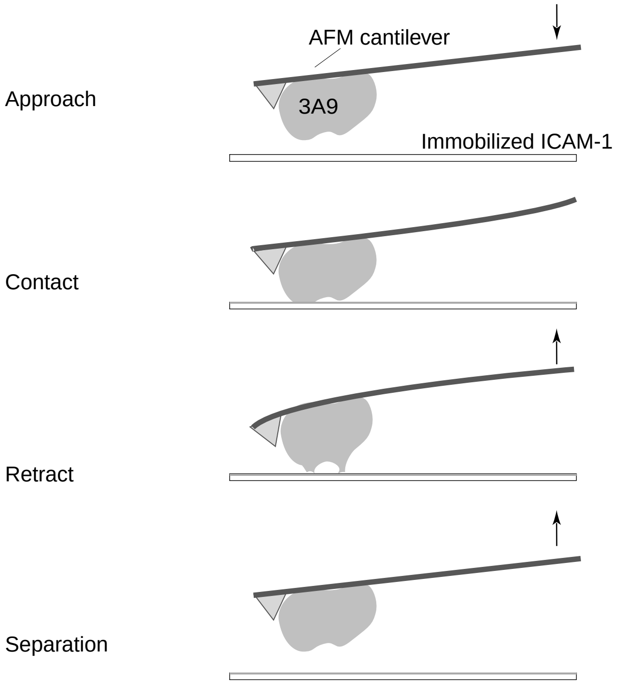


Figure 7. Steps in the acquisition of an AFM force measurement. The first step is the approach of the cantilever with a cell bound to the substrate, followed by contact between the cell and substrate and retraction of the cantilever, which results in the separation of the cell from the substrate. The cantilever is bent during this process. The arrows indicate the direction of cantilever movement.

in acetone for 5 min, UV irradiated for 30 min, and incubated in biotinamidocaproyl-labeled bovine serum albumin (biotin-BSA, 0.5 mg/ml in 100 mM $NaHCO_3$, pH 8.6; Sigma) overnight. The cantilevers are then rinsed three times with phosphate-buffered saline (PBS, 10 mM PO_4^{3-}, 150 mM NaCl, pH 7.3) and incubated in streptavidin (0.5 mg/ml in PBS; Pierce, Rockford, IL) for 10 min at room temperature. Following the removal of unbound streptavidin, the cantilevers are incubated in biotinylated Con A (0.2 mg/ml in PBS; Sigma) and then rinsed with PBS.

The actual process of attaching a cell to a cantilever tip is reminiscent of fishing. A cantilever tip is positioned above the center of the cell. The largest triangular cantilever (320 μm long and 22 μm wide) with a spring constant of 0.017 N/m on the cantilever chip is usually used in our measurements. The cell is brought into focus, with the cantilever slightly out of focus. Then, the tip is lowered onto the center of the cell and held there motionless for approximately 1 s. When attached, the cell is positioned right behind the AFM cantilever tip. The force required to dislodge the cell from the tip is greater than 2 nN, which is much greater than the forces measured in the receptor-ligand studies that were, at most, 300 pN (15).

A piezoelectric translator is used during measurement acquisition to lower the cantilever with an attached cell onto the sample. The interaction between the attached cell and the sample is given by the deflection of the cantilever. This deflection is measured by reflecting a laser beam off the cantilever into a position sensitive two-segment photodiode detector, as described in the instrumentation section above.

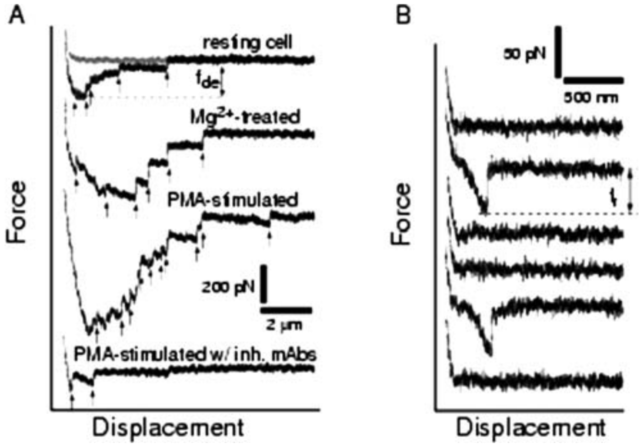


Figure 8. AFM force versus displacement traces of the interaction between cells expressing LFA-1 and immobilized ICAM-1. (a) Multiple-bond measurements acquired with a compression force of 200 pN, 5 s contact, and a cantilever retraction speed of 2 μm/s. The measurements were carried out with a resting, untreated cell (1st trace), a Mg^{2+}-treated cell (2nd trace), and a PMA-stimulated cell (3rd trace). The 4th trace corresponds to a measurement acquired from a PMA-stimulated cell in the presence of LFA-1 (20 μg/ml FD441.8) and ICAM-1 (20 μg/ml BE29G1) function-blocking monoclonal antibodies (mAbs). Arrows point to breakage of LFA-1/ICAM-1 bond(s). f_{de} is the detachment force, and the shaded area estimates the work of de-adhesion. (b) Single-molecule measurements of LFA-1/ICAM-1 unbinding forces. Traces 2 and 5 show adhesion. Measurements were obtained under conditions that minimized contact between the LFA-1-expressing cell and the ICAM-1-coated surface. The compression force was reduced to ~60 pN and the contact time to 50 ms. An adhesion frequency of less than 30% in the force measurements ensured that a >85% probability exists that the adhesion event is mediated by a single LFA-1/ICAM-1 complex (42,43). The frequency of adhesion in test and control experiments was examined to confirm the specificity of the interaction. The addition of monoclonal antibodies against either LFA-1 or ICAM-1 significantly lowered the frequency of adhesion of both resting cells and activated cells under identical experimental conditions.

As a result of this process, a force scan is obtained. The studies shown in Fig. 8 (42,43) were conducted on cells expressing the adhesion receptor LFA-1 (leukocyte function-associated antigen-1), an integrin expressed on the surface of T-cells, and its ligand ICAM-1 (intercellular adhesion molecule-1), expressed on the surface of APCs. In these experiments, LFA-1-expressing cells and ICAM-1 protein were used. An example of a few force scans from multiple bond cell adhesion studies can be seen in Fig. 8a. The red trace is the approach trace and the black is the retract trace. As the cantilever is lowered and contact is made between the cell and substrate, an initial increase in force occurs. As the cantilever is retracted back up, the force returns to zero and becomes negative as bonds are stretched and begin to break. The jumps in the force scan, which are pointed out by the arrows, represent bonds breaking. Two parameters can be used in such measurements to assess the level of cell adhesion. One is the detachment force, which is the maximum force required to dislodge the cell. Another is the work of deadhesion, which is the amount of work required to pull and stretch the cell and for the bonds to break. It is derived by integrating the adhesive force over the distance traveled by the cantilever. In this example, Mg^{2+} and PMA are used, which enhance the adhesion of the cells studied through various mechanisms. It is easily observed that a very pronounced increase occurs in the area under the curve as well as the number of bonds that were broken following the application of these adhesion stimulators (16).

FM Force Measurements of Individual Receptor/Ligand Complexes: A different set of information can be derived from single-molecule adhesion measurements. These type of studies offer insight into the dissociation pathway of a receptor-ligand interaction and the changes in free energy that are associated with this process, which is achieved by measuring single-bond interactions between a receptor and ligand at increasing loading rates (20 pN/s–50,000 pN/s). In the author's setup, it translates to using rates of retraction of the cantilever from 0.1 to 15 μm/s.

In order to obtain unbinding forces between a single receptor-ligand pair, the experiments have to be carried out in conditions that minimize contact between the cantilever tip and substrate. A > 85% probability exists that the adhesion event is mediated by a single bond if the frequency of adhesion is maintained below 30% (15,42). An example of such measurements can be seen in Fig. 8b.

Depending on the speed at which the cantilever is retracted during the measurements, the collected data usually needs to be corrected for hydrodynamic drag, which is due to the hydrodynamic force that acts in the opposite direction of cantilever movement, and its magnitude is proportional to the cantilever movement speeds. The hydrodynamic force may be determined based on the method used by Tees et al. and Evans et al. (42,43). In single-bond AFM studies, it is found that the data obtained with cantilever retraction speeds higher than 1 μm/s needed to be corrected by adding the hydrodynamic force.

The damping coefficient can be determined by plotting the hydrodynamic force versus the speed of cantilever movement. The damping coefficient is the slope of the linear fit and was found to be about 2 pN □s/μm in the author's work (15).

AFM Measurements of Cell Elasticity. The AFM can also be used as a microindenter that probes the mechanical properties of the cell. In these measurements, which enable assesment of cell elasticity, a bare AFM tip is lowered onto the center of the cell surface at a set rate, typically 2 μm/s. Following contact, the AFM tip exerts a force against the cell that is proportional to the deflection of the cantilever. The indentation force used is below 1 nN (~600 pN) in order to prevent damage to the cell. The deflection of the cantilever is recorded as a function of the piezoelectric translator position during the approach and withdrawal of the AFM tip. The force-indentation curves of the cells are derived from these records using the surface of the tissue culture dish to calibrate the deflection of the cantilever. Then, one can estimate the Young's modulus, which is a measure of elasticity. Estimates of Young's modulus are made on the assumptions that the cell is an isotropic elastic solid and the AFM tip is a rigid cone (44–46). According to this model, initially proposed by Hertz, the force (F) indentation (α) relation is a function of Young's modulus of the cell, K, and the angle formed by the indenter and the plane of the surface, θ, as follows:

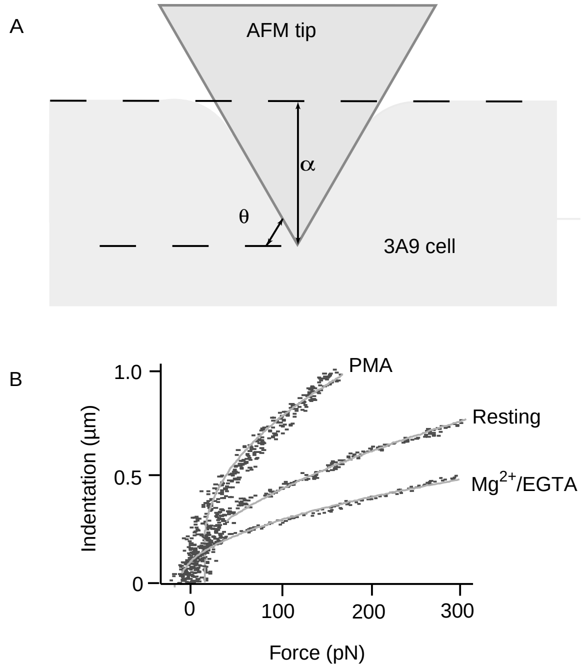


Figure 9. Acquisition of cell compliance measurements. (a) Tip of the AFM cantilever indenting a 3A9 cell. The cell compliance measurements were based on the assumption that the cell is an isotropic elastic solid and the AFM tip is a rigid cone (44–46). According to this model, initially proposed by Love and Hertz, the force (F)-indentation (α) relation (shown) is a function of Young's modulus of the cell, K, and the angle formed by the indenter and the plane of the surface, θ, as in equation (1). The indenter angle, θ, is assumed formed by the AFM tip and the 3A9 cell to be 55° and the Poisson ratio ν to be 0.5. (b) Force versus indentation traces of resting, PMA-stimulated and Mg^{2+}-treated 3A9 cells.

$$F = \frac{K}{2(1-\upsilon^2)}\,\frac{4}{\pi \tan\theta}\,\alpha^2 \qquad (1)$$

Young's modulus are obtained in the author's laboratory by least square analysis of the forceindentation curve using routines in the Igor Pro (WaveMetrics, Inc., Lake Oswego, OR) software package. The indenter angle θ and Poisson ratio ν are assumed to be 55° and 0.5, respectively.

In order to determine the cell's elasticity, the force versus indentation measurements are fitted to the curves of the Hertz model. Figure 9 (44–46) illustrates an example of such measurements acquired on cells of varying elasticity. Cells with the greatest degree of indentation at a particular applied force will have the lowest Young's modulus values and will therefore be the "softest."

Protein Folding/Unfolding. The AFM can also be used to study protein unfolding. A cantilever tip is used to pick up proteins attached to a surface, which is followed by retrac-

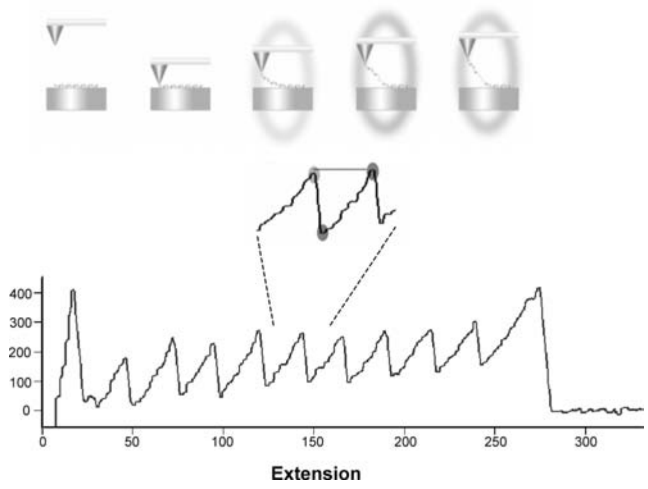


Figure 10. Consecutive unfolding peaks of a titin polyprotein, composed of FNIII domains. The inset demonstrates the corresponding steps of the unfolding process in correlation to the AFM data.

tion of the cantilever, which results in protein unfolding. The length of the unfolded protein can be over ten times its folded length, depending on the protein being studied (9).

This forced protein unfolding generates an opposing force due to the sudden drop in entropy as the protein is unfolded. Although a lower force is required to begin unfolding the protein, the force required to continue the unfolding is increased rapidly as the protein approaches its full, unfolded length. This phenomenon has been described by the worm-like chain model (WLC) of elasticity. The WLC model is based on two parameters, the total or contour length of the polymer being stretched and the persistence length. The persistence length reflects the polymer flexibility and is the length attained when a polymer is bent. A smaller persistence length is an indication of higher entropy and a polymer that is more difficult to unfold (47). When a multidomain protein is extended using the AFM, the first domain is unfolded at a certain pulling force, followed by a return of the cantilever to zero. Further unfolding meets with resistance once again, resulting in a characteristic saw-tooth profile of the unfolding, with each domain that was unfolded being represented by a peak. Figure 10 from Andreas F. Oberhauser illustrates this process. It is the unfolding of a titin polyprotein, which is composed of FNIII domains (22). The protein can also be refolded following unfolding, which is done by bringing the cantilever back down to the substrate and once again retracting it. If force curves representative of unfolding are observed once again, then refolding most likely took place. It is a much slower process (on the order of seconds) than the forced unfolding (48).

EVALUATION

Imaging AFM

The AFM is an exciting novel technology that enables the study of biological structures under physiological conditions. The AFM is probably the only technique of its kind that enables image dynamic processes taking place in real time. A number of other techniques are currently available in the biological sciences for imaging studies, however, most result in modifications to the biological sample. One such technique is electron microscopy (EM), which, until recently, provided images of the highest resolutions. In recent years, a number of modifications to the AFM have brought the resolution up to par and even surpassed those of EM.

In recent years, many advances have been made in the field of AFM. Significant improvements in resolution have been gained through cantilever tip modification. The currently available cantilevers are relatively “soft” and flexible with spring constants of 0.01–0.5 N/m. Tip deformation is one aspect that limits resolution. Recently, stiffer cantilevers have been designed improving resolution. One example are quartz cantilevers with spring constants on the order of 1 kN/m allowing for possibly subatomic-level resolution (34). Smaller cantilevers have been designed that increase the possible scanning speed of the AFM. Images of 100×100 pixels (240 nm scan size) have been achieved in 80 ms. A sharper, finer tip can also improve resolution, which has been achieved through the use of carbonanotubes, probably the greatest probe improvement to date, which are seamless cylinders composed of sp^2-bonded Carbon (49).

Several characteristics exist that make Carbon nanotubes improved AFM tip materials, including small diameter, a high aspect ratio, large Young’s modulus, and mechanical robustness. They are able to elastically buckle under large loads. All these properties translate into higher sample resolution. Chemical vapor deposition has made it easier to grow Carbon nanaotubes on cantilever surfaces, a process that replaces previously more laborious and time-consuming attachment techniques (4,50).

A few techniques also exist worth mentioning that have improved AFM imaging. One such method is cryoAFM. This method addresses the previously mentioned problem of tip and sample flexibility. In this case, samples are imaged at extremely cold temperatures in their cryogenic states, which provides a rigid surface that exhibits a high Young’s modulus, thus reducing cantilever deformation and increasing resolution. CryoAFM images match and surpass EM images. Other improvements address the problem resulting from the vibration induced by the commonly used piezoelectric translator. This vibration is translated to the cantilever holder and the liquid containing the sample being imaged. Magnetic mode (MAC) eliminates the cantilever holder entirely and replaces it with a magnetic cantilever. The cantilever is manipulated via a magnetic field. Photothermal mode (PMOD) uses a bimetallic cantilever that is oscillated via a pulsed diode laser (50,51).

Advances have also been made in single-molecule manipulation with a nanomanipulator. This method relies on a force feedback pen that actually allows the user to touch and manipulate the sample being studied. For example, one can dissect DNA from a chromosome. The interaction forces involved during manipulation of samples can also be studied through oscillating mode imaging. The nanomanipulator can now measure forces in the pN–μN range. For excellent reviews on this technique, see Yang et al. (50) Fotiadis et al. (52).

Progress has also been made in imaging of membrane proteins, which are not ideal candidates for X-ray crystallography, as they do not readily form 3D crystals. Atomic-level resolution images have been obtained of membrane proteins complexed with lipids using EM. However, AFM images of these proteins offer an improvement in that they can be carried out in near physiological conditions and allow for the acquisition of functional and structural information (50,52).

The continued improvements leading to the enhanced imaging capabilities of AFM are reflected in the most recent work being done in the field. We would like to highlight one area where a great deal of progress has been made, which is the filed of DNA and RNA assembly of nanostructures in which the imaging AFM plays a pivotal role. Some of the earlier successes in this area included the construction of 2D DNA arrays, which were assembled in a predictable manner (53). Much progress has also been made with RNA in an attempt to design self-assembling building blocks. The goal of such studies is to generate molecular materials, the geometry of which can be determined for applications in nanobiotechnology (54–56). Chworos et al. were able to form stable RNA structures termed "tectosquares" from RNA helices without the presence of proteins (5). "TectoRNAs" can be thought of as RNA Lego blocks that can be used for the formation of supramolecular structures. In order for these structures to be assembled, the right conditions have to be met in terms of divalent ion concentrations, temperature, as well as the strength, length, and orientation of the RNA. The AFM is an essential tool used in this process as it allows the researcher to obtain detailed images of the assembled tectosquares providing the opportunity to compare predicted structures with those that actually formed. The structures formed by Chworos et al. were in good agreement with the predicted structures. Figure 11 demonstrates the amazing predictability of assembly of these structures, where nine different types of RNA patterns were created (5). Such DNA and RNA structures may have applications in nanotechnology and the material sciences as they could be used to generate nanochips, nanocircuits, and nanocrystals (57). For an excellent recent review of DNA nanomechanical devices, please see Seeman (58).

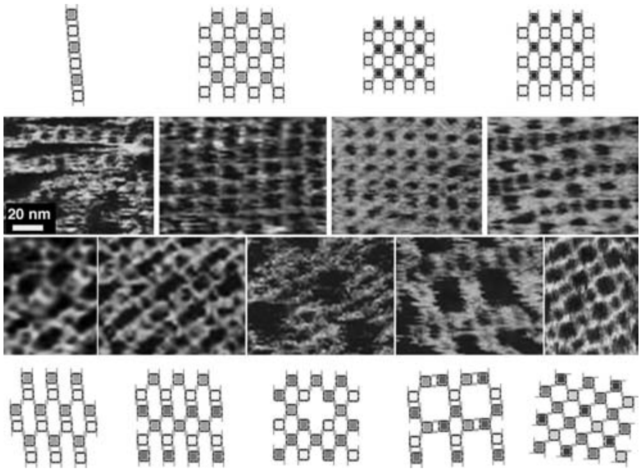


Figure 11. Diagram and AFM images of tectosquare nanopatterns generated from 22 tectosquares. One micrometer square scale AFM images obtained in solution (clockwise from the upper leftmost image) for the ladder pattern, fish net pattern, striped velvet pattern, basket weave pattern, cross pattern, tartan pattern, polka dot pattern, lace pattern, and diamond pattern. Scale bar, 20 nm.

Force Spectroscopy

Force spectroscopy allows us to measure interaction forces between receptors and their respective ligands. These studies can be carried out with purified proteins or cells, or a combination of both. Traditionally, adhesion measurements have been conducted using adhesion assays, which involve the attachment of cells to dishes coated with substrate. The cells are later dislodged manually or using a centrifuge, which are older yet still viable techniques that provide basic kinetic information about the interaction of a receptor-ligand pair. More advanced techniques for conducting force measurements include the use of microneedles, optical tweezers, magnetic beads, and the biomembrane force probe. These techniques, much like the AFM, provide more advanced information from which energy landscapes of an interacting receptor-ligand pair may be determined (11).

The AFM is also a powerful tool for determining the mechanical properties of cells, which was traditionally done using micropipettes or the cell poker, which offered much less precision than the AFM. More recently, methods such as the scanning acoustic microscope, optical tweezers, and magnetic tweezers have also been used in addition to the AFM (59).

An important advantage of the AFM over other methods is that it can be used in conjunction with other techniques through relatively simple modifications. Recently, it has been combined with the patch clamp technique to study the mechanically activated ion channels of sensory cells of the inner ear. This strategy allowed the researchers to use the AFM tip to stimulate the mechanosensory hair bundles by exerting force on them and measure the electrical output of the patch clamp simultaneously (9,60). Another example is combining an AFM with a confocal microscope, which could allow one to monitor cellular responses to AFM measurements using fluorescent reporter systems. One could monitor calcium levels, expression of caspases, and so on (61,62). The AFM could also be combined with RICM microsopy as well as FRET.

Other recent advances involve modifications that would allow for more efficient and effective receptor-ligand studies, including the use of more than one cantilever on the same chip simultaneously. In this case, multiple proteins could be attached and their interaction with their ligand could be measured. So far, this approach has been done with two cantilevers, which involves the use of two laser beams. Further modifications could allow for measurements with even more proteins. Improvements can also be mad in plating of the ligands proteins. In the ideal scenario, different proteins could be plated so that interaction between different receptor-ligand pairs cools be carried out simultaneously. Current improvements also involve finding better ways to attach cells and proteins to the cantilever that would result in covalent attachment to the tip. Another area that requires improvement is data analysis. The currently available analysis program involves days of rather tedious computer time. Automating analysis would greatly reduce the time required to study a particular interaction. Also, some improvements can be made in data acquisition, where still frequent adjustments of cantilever retraction speed and contact time are required throughout the course of the experiment. Automating data acquisition would allow experiments to be carried out during the night, when noise levels are also minimal.

The applications of AFM technology are vast and too numerous to describe in one review article. The author's have attempted to summarize the technology that was deemed to be of great importance in the developing field of AFM. AFM technology is still limited to a relatively small number of laboratories, which is most likely due to the lack of familiarity with the field, limited expertise in operation, as well as the expense involved in acquiring an AFM. However, it is changing as more and more people are discovering the possibilities that become open to them if they acquire and familiarize themselves with this technology.

BIBLIOGRAPHY

1. Binnig G, Quate CF, Gerber C. Atomic force microscope. Phys Rev Lett 1986;56:930–933.
2. Heinz WF, Hoh JH. Spatially resolved force spectroscopy of biological surfaces using the atomic force microscope. Trends Biotechnol 1999;17(4):143–150.
3. Engel A, Lyubchenko Y, Muller D. Atomic force microscopy: A powerful tool to observe biomolecules at work. Trends Cell Biol 1999;9(2):77–80.
4. Hansma HG. Surface biology of DNA by atomic force microscopy. Annu Rev Phys Chem 2001;52:71–92.
5. Chworos A, et al. Building programmable jigsaw puzzles with RNA. Science 2004;306(5704):2068–2072.
6. Kasas S, et al. Escherichia coli RNA polymerase activity observed using atomic force microscopy. Biochemistry 1997; 36(3):461–468.
7. Rivetti C, et al. Visualizing RNA extrusion and DNA wrapping in transcription elongation complexes of bacterial and eukaryotic RNA polymerases. J Mol Biol 2003;326(5):1413–1426.
8. Pesen D, Hoh JH. Modes of remodeling in the cortical cytoskeleton of vascular endothelial cells. FEBS Lett 2005;579(2): 473–476.
9. Horber JKH. Local probe techniques. Methods Cell Biol 2002;68:1–32.
10. Lee GU, Kidwell DA, Colton RJ. Sensing discrete streptavidin-biotin interactions with AFM. Langmuir 1994;10(2):354–361.
11. Zhang X, Chen A, Wojcikiewicz E, Moy VT. Probing ligand-receptor interactions with atomic force microscopy. In: Protein-Protein Interactions: A Molecular Cloning Manual. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 2002. p 241–254.
12. Yuan C, et al. Energy landscape of streptavidin-biotin complexes measured by atomic force microscopy. Biochemistry 2000;39(33):10219–10223.
13. Benoit M, et al. Discrete interactions in cell adhesion measured by single molecule force spectroscopy. Nat Cell Biol 2000;2(6):313–317.
14. Benoit M. Cell adhesion measured by force spectroscopy on living cells. Methods Cell Biol 2002;68:91–114.
15. Zhang X, Wojcikiewicz E, Moy VT. Force spectroscopy of the leukocyte function-associated antigen-1/intercellular adhesion molecule-1 interaction. Biophys J 2002;83(4):2270–2279.
16. Wojcikiewicz EP, et al. Contributions of molecular binding events and cellular compliance to the modulation of leukocyte adhesion. J Cell Sci 2003;116(12):2531–2539.
17. Matzke R, Jacobson K, Radmacher M. Direct, high-resolution measurement of furrow stiffening during division of adherent cells. Nat Cell Biol 2001;3(6):607–610.
18. Hassan AE, et al. Relative microelastic mapping of living cells by atomic force microscopy. Biophys J 1998;74(3):1564–1578.
19. Wojcikiewicz EP, Zhang X, Moy VT. Force and compliance measurements on living cells using atomic force microscopy (AFM). Biol Proced Online 2004;6:1–9.
20. Rief M, et al. Reversible unfolding of individual titin immunoglobulin domains by AFM [see comments]. Science 1997;276(5315):1109–1112.
21. Oesterhelt F, et al. Unfolding pathways of individual bacteriorhodopsins [see comments]. Science 2000;288(5463):143–146.
22. Li H, et al. Reverse engineering of the giant muscle protein titin. Nature 2002;418(6901):998–1002.
23. Sarid D. Scanning Force Microscopy. New York: Oxford University Press; 1991.
24. Meyer G, Amer NM. Novel optical approach to AFM. Appl Phys Lett 1988;53:1045–1047.
25. Israelachvili JN. Intermolecular and Surface Forces. 2nd ed. London: Academic Press; 1992.
26. Meyer G, Amer NM. Simultaneous measurement of lateral and normal forces with an optical-beam-deflection AFM. Appl Phys Lett 1990;57(20):2089–2091.
27. Mate CM, et al. Atomic-scale friction of a tungsten tip on a graphite surface. Phys Rev Lett 1987;59(17):1942–1945.
28. Overney RM, et al. Friction measurements on phasesegregated thin films with a modified atomic force microscope. Nature 1992;359:133–135.
29. Frisbie CD, et al. Functional group imaging by chemical force microscopy. Science 1994;265:2071–2074.
30. Piner RD, et al. "Dip-Pen" nanolithography. Science 1999; 283(5402):661–663.
31. Kwak KJ, et al. Topographic effects on adhesive force mapping of stretched DNA molecules by pulsed-force-mode atomic force microscopy. Ultramicroscopy 2004;100(3–4):179–186.
32. Magonov SN, EV, Whango MH. Phase imaging and stiffness in tapping mode AFM. Surf Sci 1997;375:L385–L391.
33. Miyatani T, Horii M, Rosa A, Fujihira M, Marti O. Mapping of electrical double-layer force between tip and sample surfaces in water with pulsed- forcemode atomic force microscopy. Appl Phys Lett 1997;71:2632–2634.
34. Giessibl FJ, et al. Subatomic features on the silicon (111)-(7 × 7) surface observed by atomic force microscopy. science 2000;289(5478):422–426.
35. Morita SR, Wiesendanger R, Meyer E. Noncontact AFM. New York: Springer; 2002.
36. Hembacher S, Giessibl FJ, MJ. Force microscopy with light-atom probes. Science 2004;305(5682):380–383.
37. Sader JE. Parallel beam approximation for V-shaped atomic force micrscope cantilevers. Rev Sci Instrum 1995;66:4583–4587.
38. Senden TJ, Ducker WA. Experimental determination of spring constants in atomic force microscopy. Langmuir 1994;10: 1003–1004.

39. Hutter JL, Bechhoefer J. Calibration of atomic-force microscope tips. Rev Sci Instrum 1993;64(7):1868–1873.
40. Bell GI. Models for the specific adhesion of cells to cells. Science 1978;200:618–627.
41. Merkel R, et al. Energy landscapes of receptor-ligand bonds explored with dynamic force spectroscopy [see comments]. Nature 1999;397(6714):50–53.
42. Tees DFJ, Woodward JT, Hammer DA. Reliability theory for receptorligand bond dissociation. J Chem Phys 2001;114: 7483–7496.
43. Evans E. Probing the relation between force—lifetime—and chemistry in single molecular bonds. Ann Rev Biophys Biomolec Struc 2001;30:105–128.
44. Hoh JH, Schoenenberger CA. Surface morphology and mechanical properties of MDCK monolayers by atomic force microscopy. J Cell Sci 1994;107:1105–1114.
45. Radmacher M, et al. Measuring the viscoelastic properties of human platelets with the atomic force microscope. Biophys J 1996;70(1):556–567.
46. Wu HW, Kuhn T, Moy VT. Mechanical properties of L929 cells measured by atomic force microscopy: Effects of anticytoskeletal drugs and membrane crosslinking. Scanning 1998;20(5): 389–397.
47. Fisher TE, et al. The study of protein mechanics with the atomic force microscope. Trends Biochem Sci 1999;24(10):379–384.
48. Altmann SM, Lenne P-F. Forced unfolding of single proteins. Methods Cell Biol 2002;68:312–336.
49. Hafner JH, et al. Structural and functional imaging with carbon nanotube AFM probes. Prog Biophys Molec Biol 2001;77(1):73–110.
50. Yang Y, Wang H, Erie DA. Quantitative characterization of biomolecular assemblies and interactions using atomic force microscopy. Methods 2003;29(2):175–187.
51. Sheng S, Zhifeng S. Cryo-atomic force microscopy. Methods Cell Biol 2002;68:243–256.
52. Fotiadis D, et al. Imaging and manipulation of biological structures with the AFM. Micron 2002;33(4):385–397.
53. Winfree E, et al. Design and self-assembly of two-dimensional DNA crystals. Nature 1998;394(6693):539–544.
54. Hansma HG, Kasuya K, Oroudjev E. Atomic force microscopy imaging and pulling of nucleic acids. Curr Opin Struct Biol 2004;14(3):380–385.
55. Jaeger L, Westhof E, Leontis NB. TectoRNA: Modular assembly units for the construction of RNA nano-objects. Nucl Acids Res 2001;29:455–463.
56. Seeman NC. DNA in a material world. Nature 2003;421:427–431.
57. Yan H, et al. DNA-templated self-assembly of protein arrays and highly conductive nanowires. Science 2003;301(5641): 1882–1884.
58. Seeman NC. From genes to machines: DNA nanomechanical devices. Trends Biochem Sci 2005;30(3):119–125.
59. Radmacher M. Measuring the elastic properties of living cells by AFM. Methods Cell Biol 2002;68:67–90.
60. Langer MG, Koitschev A. The biophysics of sensory cells of the inner ear examined by AFM and patch clamp. Methods Cell Biol 2002;68:142–171.
61. Charras GT, Lehenkari P, Horton M. Biotechnological applications of AFM. Methods Cell Biol 2002; 68.
62. Charras GT, Horton MA. Single cell mechanotransduction and its modulation analyzed by atomic force microscope indentation. Biophys J 2002;82(6):2970–2981.

See also BIOMAGNETISM; NANOPARTICLES.

MICROSCOPY, SCANNING TUNNELING

VIRGINIA M. AYRES
Michigan State University
East Lansing, Michigan

INTRODUCTION

Four years after its invention in 1982 (1), the scanning tunneling microscope (STM) was awarded the 1986 Nobel Prize for physics, one of only four such prestigious awards given for a truly significant contribution to scientific instrumentation. Since then, the family of scanning probe microscopy (SPM) techniques, which includes scanning tunneling microscopy, atomic force microscopy (2–4), magnetic force microscopy (5), near-field optical microscopy (6), scanning thermal microscopy (7), and others, has revolutionized studies of semiconductors, polymers, and biological systems. The key capability of SPM is that, through a controlled combination of feedback loops and detectors with the raster motion of piezoelectric actuator, it enables direct investigations of atomic-to-nanometer scale phenomena.

Scanning probe microscopy is based on a piezoelectric-actuated relative motion of a tip versus sample surface, while both are held in a near-field relationship with each other. In standard SPM imaging, some type of tip-sample interaction (e.g., tunneling current, Coulombic forces, magnetic field strength) is held constant in z through the use of feedback loops, while the tip relative to the sample undergoes an x–y raster motion, thereby creating a surface map of the interaction. The scan rate of the x–y raster motion per line is on the order of seconds while the tip-sample interaction is on the order of nanoseconds or less. The SPM is inherently cable of producing surface maps with atomic scale resolution, although convolution of tip and sample artifacts must be considered.

Scanning tunneling microscopy is based on a tunneling current from filled to empty electronic states. The selectivity induced by conservation of energy and momentum requirements results in a self-selective interaction that gives STM the highest resolution of all scanning probe techniques. Even with artifacts, STM routinely produces atomic scale (angstrom) resolution.

With such resolution possible, it would be highly desirable to apply STM to investigations of molecular biology and medicine. Key issues in biology and medicine revolve around regulatory signaling cascades that are triggered through the interaction of specific macromolecules with specific surface sites. These are well within the inherent resolution range of STM.

The difficulty when considering the application of STM to molecular biology is that biological samples are non-conductive. It may be more accurate to describe biological samples as having both local and varying conductivities. These two issues will addressed in this article, and examples of conditions for the successful use of STM for biomedical imaging will be discussed. We begin with an overview of successful applications of STM in biology and medicine.

SCANNING TUNNELING MICROSCOPY IN BIOLOGY AND MEDICINE: DNA AND RNA

The STM imaging for direct analysis of base pair arrangements in DNA was historically the first biological application of the new technique. An amusing piece of scientific history is that the first (and widely publicized) images (8–12) of (deoxyribonucleic acid) DNA were subsequently shown to correspond to electronic sites on the underlying graphite substrate! However, more careful investigations have resulted in an authentic body of work in which the base pairings and conformations of DNA and RNA are directly investigated by STM. One goal of these investigations is to replace bulk sequencing techniques and crystal diffraction techniques, which both require large amounts of material, with the direct sequencing of single molecules of DNA and RNA. Two examples of DNA and RNA investigation by STM are presented here. One is an investigation of DNA and RNA structures, and the other is an investigation of DNA biomedical function.

Recently reported research from the group at The Institute for Scientific and Industrial Research at Osaka University in Japan (13) has shown detailed STM images of well-defined guanine-cytosine (G-C) and adenine-thymine (A-T) base pairings in double- and single-stranded DNA. Four simple samples involving only G-C and only A-T base pairs in mixed (hetero) and single sided (homo) combinations were chosen for analysis (Fig. 1). These were deposited on a single-crystal copper (111)-orientation [Cu(111)] substrate using a technique developed specially by this group to produce flat, extended strands for imaging. An STM image showing the individual A-T base pairs in the hetero A-T sample is shown in Fig. 2. Images of the overall structures indicated repeat distances consistent with interpretation as the double helix. Images from mixed samples of hetero G-C and hetero A-T are shown in Fig. 3. The larger structure is interpreted as hetero G-C and the smaller as hetero A-T, which is consistent with X-ray diffraction data that indicates the A-T combination is more compact.

Only the double helix structure was observed for the hetero G-C samples. However, the homo G-C structures, hetero A-T structures, and homo A-T structures were observed in two types, and the spot spacings and sizes of the second type would be consistent with interpretation as single-stranded DNA. The observed presence or lack of single-stranded configurations among the samples is consistent with the fact that hetero G-C has a higher melting (unraveling) temperature than the homo G-C and thus is more difficult to unwind. Both hetero and homo A-T pairs have lower melting temperatures than either of the G-C pairs. Images of both hetero A-T and Homo A-T samples often showed sizing and spacings consistent with interpretation as single-stranded DNA, in addition to observed double helix specimens. Thus, the presence/lack of single-stranded versus double helix images is consistent with known melting temperature data for the C-G and A-T base pairings.

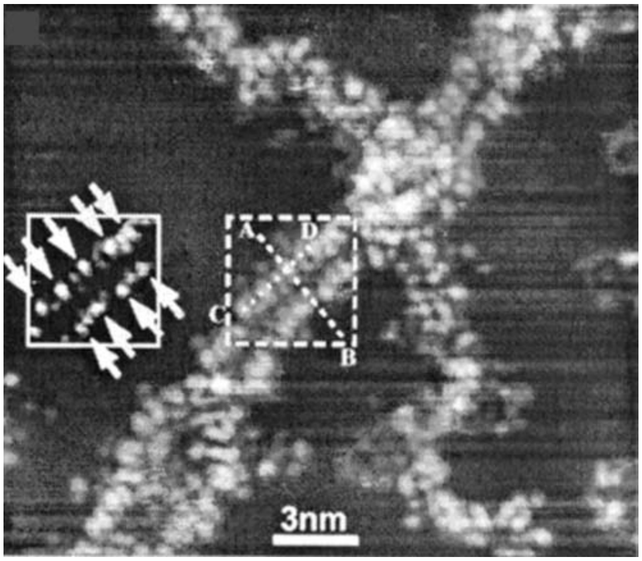


Figure 2. STM image of portion of Hetero A-T double helix of showing base pairs. (Reproduced from Ref. 9, used with permission.)

The same group has also reported successful STM investigations of transfer-ribonuclic acid (t-RNA) (14). In

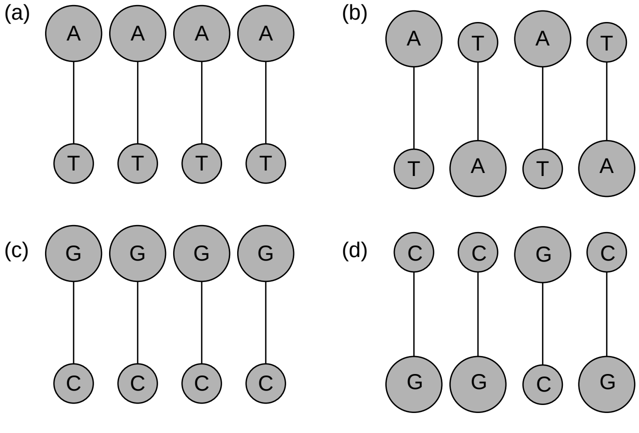


Figure 1. (a) Homo A-T, (b) Hetero A-T, (c) Homo G-C, and (d) Hetero G-C. (Figure adapted from Ref. 9, used with permission.)

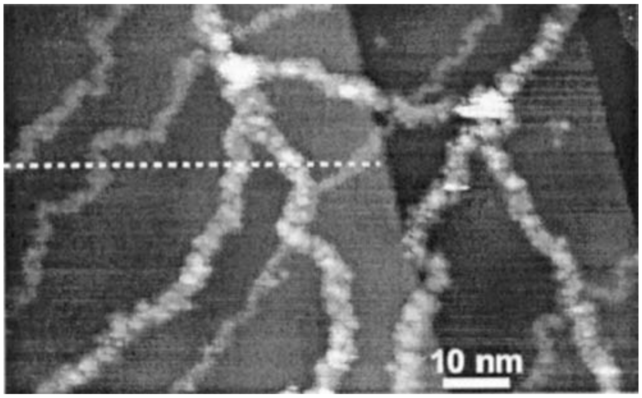


Figure 3. Hetero G-C and Hetero A-T mixed sample. The larger specimens are identified as Hetero G-C, and the smaller specimens are identified as Hetero A-T. Both are in a double helix configuration. (Reproduced from Ref. 9, used with permission.)

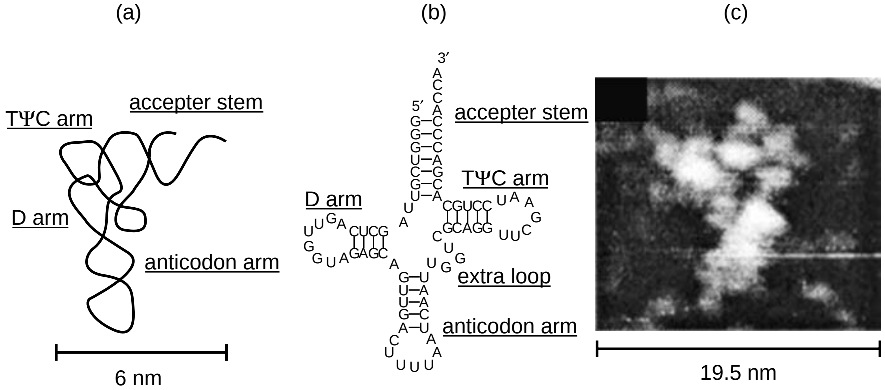


Figure 4. (a) Model of t-RNA L-shaped conformation. (b) Model base pair arrangement in L-shaped conformation. (c) STM image of L-shaped conformation at physiological pH. (Reproduced from Ref. 10, used with permission.)

RNA, the base pairing is adenine-uracil (A-U) instead of adenine-thymine (A-T). Also the backbone sugars are ribose rather than deoxyribose, but are still linked by phosphate groups. The RNA is very difficult to synthesize as a single crystal and consequently there is a very limited amount of X-ray diffraction data available for RNA. Little is known about its variations, and therefore direct investigations of single molecule RNA would add much to our knowledge.

Transfer RNA is a small RNA chain of ~ 74–93 nucleotides that transfers a specific amino acid to a growing polypeptide chain at the ribosomal site of protein synthesis during translation (15). It has sites for amino acid attachment, and an anticodon region for codon recognition that binds to a specific sequence on the messenger RNA (m-RNA) chain. It has a partial double-helix structure even though it has only one chain, because the single RNA chain folds back, and loops back, on itself, as shown in Fig. 4a.

X-ray diffraction studies (16) have indicated that the t-RNA structure may often assume an L-shaped conformation with a long and a short arm. A model of the Escherichia Coli lysine t-RNA macromolecule used by the group for its STM studies is shown in Fig. 4a and b. It shows both the L conformation and the underlying loop and base pair chemistry.

Using STM, the group was able to directly image the L conformation as shown in Fig. 4c. In addition to the first direct statistical data on the lengths of the long and short arms, obtained from analysis of several STM images, an analysis of the influence of pH on conformation was also carried out. Current investigations are focusing on biofunction research issues in addition to structural research issues, using STM to directly image the coupling of the important amino acid molecules at specific t-RNA sites.

The STM investigations of nanobiomedical rather than structural issues are an important emerging research area. One example is the recently reported research from the University of Sydney group in which the local binding of retinoic acid, a potent gene regulatory molecule, to plasmid p-GEM-T easy (596 base pair Promega) DNA fragments on a single-crystal graphite substrate, was directly imaged and reported (17). Retinoic acid has been documented as responsible for a number of profound effects in cell differentiation and proliferation, and is known to accomplish its functions through selective site binding during the transcription process. The STM images of retinoic acid by itself on a single-crystal graphite substrate were investigated first. These showed sizes consistent with the retinoic acid molecular structure, and a bright head area with a darker tail area. A molecular model of retinoic acid, also shown in Fig. 5a, shows its aliphatic carbon ring head and polymeric tail. For reasons further discussed below, the aliphatic ring head may be expected to have a higher tunneling current associated with it than the polymeric tail, and therefore the observed bright and dark areas are consistent with the expected structure.

At low concentrations, retinoic acid was observed to bind selectively at minor groove sites along the DNA, with some clustering of retinoic acid molecules observed, as shown in Fig. 5b. High resolution STM imaging provided direct evidence for alignment of the retinoic acid molecules head-to-tail structure edge-on with the minor groove and also in steric alignment with each other. From STM height studies, it could also be inferred that the aliphatic ring head was attached to a ring partner along the minor groove surface, but that the tail was not attached. This may suggest a loosely bound on–off functional mechanism. At high concentrations, retinoic acid was observed to bind along the whole length of the DNA double helix, but again selecting the minor grooves. These first direct studies of selective site binding of retinoic acid with the minor groove of DNA should serve as a template for further direct investigations of other known minor groove binders, thereby opening up the direct investigation of an entire

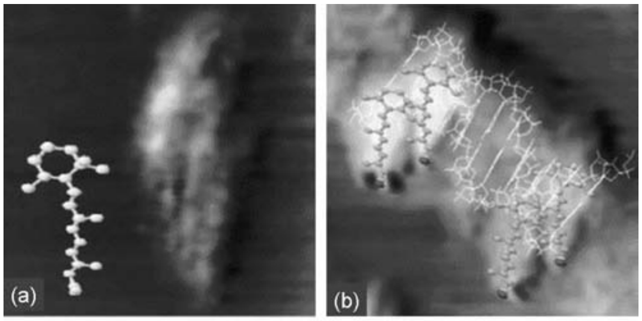


Figure 5. (a) STM image of retinoic acid on a graphite substrate compared with its molecular model showing the aliphatic ring head and polymeric tail. (b) STM image of retinoic acid binding to t-RNA with molecular model overlay. (Reproduced from Ref. 13, used with permission.)

class of regulatory molecule–DNA interactions. The interactions of related structures that are candidate therapeutic drug molecules could be receive similar direct investigation.

Note that both of the above groups have also made important contributions to sample preparation techniques for successful STM analysis of DNA and RNA. These sample preparation techniques will be discussed below in the context of the basic physics of the STM interaction, and the basic chemistry and conductivity of DNA and RNA samples.

BASIC PHYSICS OF THE STM INTERACTION

The STM is based on tip–sample interaction via a tunneling current between filled electronic states of the sample (or tip) into the empty electronic states of the tip (or sample), in response to an applied bias, as shown in Fig. 6. The bias may be positive or negative, and different and valuable information may often be obtained by investigation of the how the sample behaves in accepting, as well as in giving up, electrons. In STM imaging, it is important to recognize that the feature map or apparent topography of the acquired image is really a map of the local density of electronic states. Bright does not correspond to a raised topography; it corresponds to a region with a high density of electronic states. Therefore, in STM imaging of biological samples, an important consideration is that a differential conductivity will be observed from regions, such as rings (usually high) versus regions, such as alkane backbones (usually low). As in all SPM techniques, a z-direction feedback loop maintains some aspect of the tip samples interaction constant (Fig. 6). The readily available choices on commercial machines are to hold either the tunneling distance d constant (constant height mode) or the magnitude of the tunneling current content (constant current mode).

The current in question is a tunneling current, which is a quantum mechanical phenomenon. It is well documented that all electrons within the atomic planes of any material are in fact in such tight quarters that they display the characteristics of a wave in a waveguide, in addition to their particle-likeness. An electron at the surface of a material faces a wall (barrier) created by the dissimilar material (e.g., air, vacuum, or a liquid). While a particle would run into a wall and bounce back, a wave can penetrate into and indeed through a wall (as light goes through glass). This is illustrated in Fig. 7. Additionally, all materials have precise energy levels within them, and therefore, electrons will move by going from one energy level at location 0 to another at location d, meeting conservation of energy requirements.

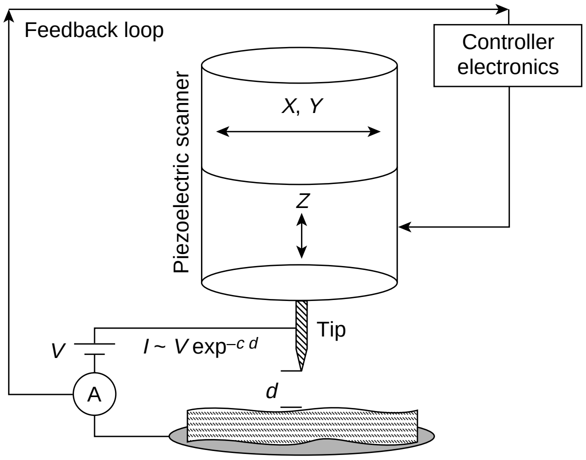


Figure 6. Important features of an STM system.

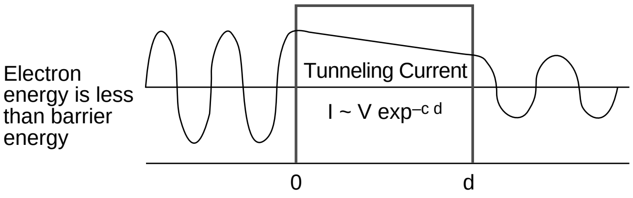


Figure 7. A particle penetrating into and through a wall.

In STM, a tip with empty electronic states is brought physically close to a sample surface. The electrons are given a direction through the application of the bias (positive in this example). Because they are wavelike, when they reach the sample surface, they can tunnel through the barrier created by the 0-to-d gap and reach the empty states of the tip, where they are recorded as a current proceeding from sample to tip. A tunneling current has the known mathematical form: $I \sim V \exp^{-cd}$, where I is the tunneling current, V is the bias voltage between the sample and the tip, c is a constant and d is the tip-sample separation distance. The tunneling current depends sensitively on the size of the 0-to-d gap distance. To observe a tunneling current, the gap must be on the order of tens of nanometers. This is the case in any commercial STM system. It is remarkable, that with the addition of a simple feedback loop, a tip can be easily maintained within nanometers of a sample surface without touching it. Typical STM tunneling currents are on the order of 10^{-9}–10^{-12} A. With special preamplifiers, currents on the order of 10^{-14} A can be detected.

Because STM is a current-based technique, some situations that can interfere with its current will be briefly discussed. Very common in STM imaging of biological samples is for the tip to acquire a layer of biological material, possibly by going too close to the sample surface while passing over an insulating region where the feedback loop has little to work on. This usually just introduces image artifacts, discussed below, but it can sometimes insulate the tip from the sample, thus terminating the tip–sample interaction. The problem can be minimized through careful consideration of the expected chemistry and local conductivity of the biological specimen to be investigated.

CHEMISTRY, CONFORMATION, AND CONDUCTIVITY OF BIOLOGICAL SAMPLES

Consideration of the basic chemistry involved in a biological sample can help to determine its appropriateness for STM imaging. The building blocks for DNA and RNA are

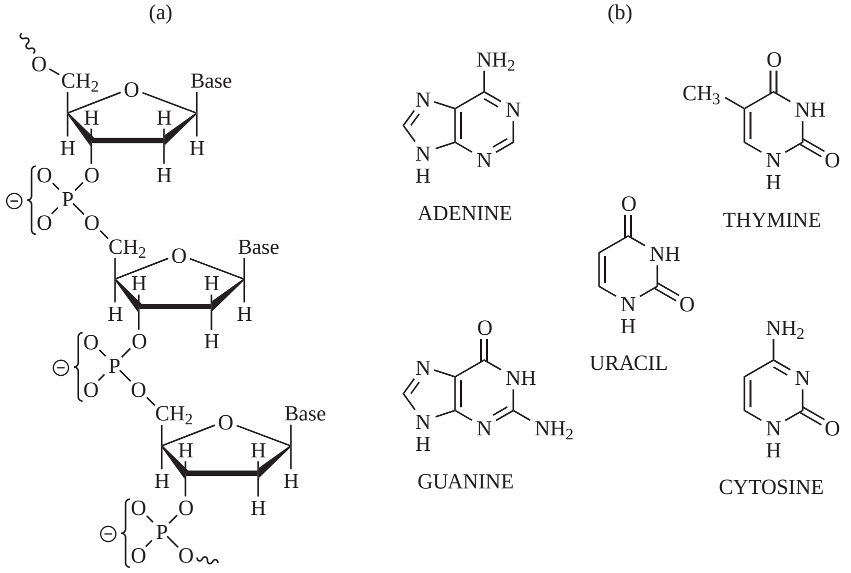


Figure 8. (a) The deoxyribose (ribose) sugar/phosphate backbone for DNA (RNA) is negatively charged due to phosphate groups. (b) DNA and RNA bases are nitrogenous ring systems.

shown Fig. 8. The sugar-phosphate backbone contains negatively charged phosphate groups for both DNA and RNA. The bases adenine, thymine, uracil, guanine, and cytosine are all nitrogenous ring systems. Thymine, cytosine, and uracil are six-member ring pyrimidine systems, and adenine and guanine are purines, the fusion of a six-member pyrimidine ring to a five-member imidazole ring. Successful STM imaging of monolayers of the individual bases has been reported (18,19). Examples of the high resolution STM imaging that is possible for monolayers of the individual bases are shown in Figs. 9 and 10.

The nitrogenous ring systems, like the classic benzene ring system, which has also been imaged (20), contain *p*-orbital electrons above and below the ring structure plane, which create a conductive electron cloud. Hence, the successful STM imaging of the DNA and RNA systems by the Osaka University and University of Sydney groups might be expected from the charged phosphate groups in the backbones and the ring systems in the base pairs.

However, there are also very difficult issues to resolve in making the local conductivity of, especially, the signature DNA and RNA base pairs available to the STM tip. These are enclosed within the sugar-phosphate backbones, and only partially exposed by the twisting of the helix, as shown in Fig. 11a and b (21,22). Also, the choice of substrate will powerfully influence the molecular structure deposited on it, especially if it is small. An example of this is shown in Fig. 12, taken from Ref. 16. The behaviors of pyridine (a single-nitrogen close relation to pyrimidine) and benzene on a single crystal (001) orientation copper, Cu(001), substrate were investigated. The pyrimidine monolayers (thymine, cytosine, and uracil) in Figs. 9 and 10 had rings oriented parallel to the substrate surface, but individual pyridine molecules on Cu(001) had rings perpendicular to the surface, due to the strong nitrogen-copper atom interaction, as shown in Fig. 12a. Also, if a single hydrogen atom was dissociated from the pyridine molecule, as can happen during routine scanning, the molecule would shift its position on the copper substrate (Fig. 12b). The STM imaging of an individual benzene molecule indicated a ring system parallel to the copper substrate (Fig. 12c), but hydrogen dissociation would cause the benzene molecule to become perpendicular to the substrate surface (Fig. 12d). Therefore both the substrate choice and interactions with the imaging tip can influence the conformation of the biomolecule and whether its locally conductive portions are positioned to produce a tunneling current.

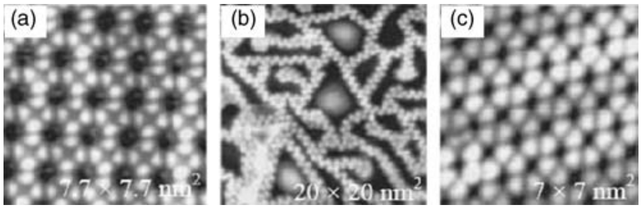


Figure 9. STM images of (a) guanine, (b) cytosine, and (c) adenine monolayers on a single crystal (111)-orientation gold substrate. (Reproduced from Ref. 14, used with permission.)

Now consider the situation of a molecule with a difference in local conductivity, like retinoic acid. The aliphatic ring head would similarly be expected to have a high local conductivity, and separate investigations of just retinoic acid by the University of Sydney group confirmed that this is the case (Fig. 5a). The polymeric tail is basically an alkane system without any *p*-type orbitals. Its conductivity is therefore expected to be less than the ring system and this is experimentally observed. However, results such as those shown in Fig. 13 from a group at California Institute of Technology, demonstrate that high resolution STM imaging even of low conductivity alkane systems is possible (23–26). Therefore, one aspect of STM biomolecular imaging is that there may be large differences in the conductivities of two closely adjacent regions. It then becomes an issue of whether the STM feedback loop will be able to sufficiently respond to the differences to maintain the tip-sample tunneling current interaction throughout the investigation. Prior consideration of the imaging parameters necessary for successful STM imaging of the *least* conductive part of the bio molecule can help.

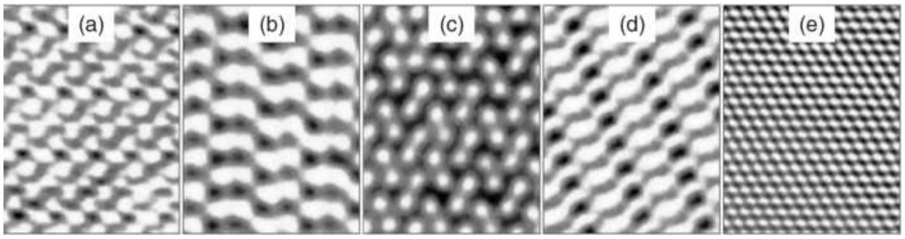


Figure 10. STM images of (a) guanine, (b) adenine, (c) uracil, and (d) thymine monolayers on (e) a single crystal (0001)-orientation molybdenum dissulfide substrate. (Adapted from Ref. 15, used with permission.)

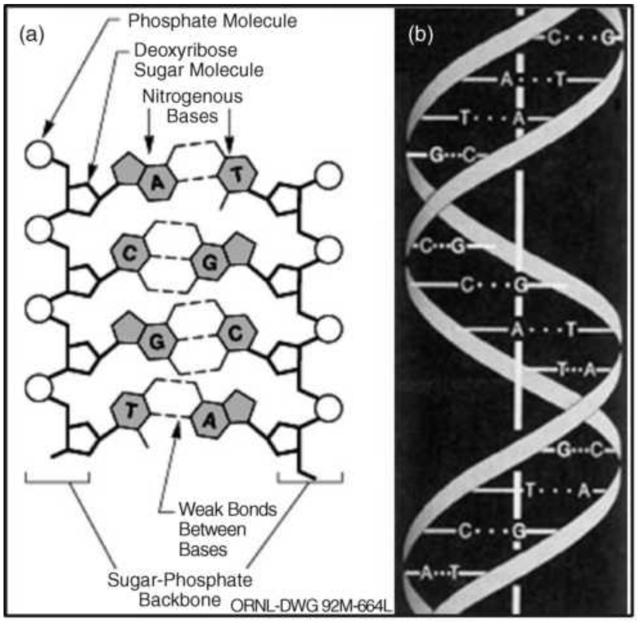


Figure 11. The three-dimensional conformation of DNA. (a) The base pairs are positioned between the sugar-phosphate backbones. (b) The overall structure is a double helix. (Reproduced from Refs. 17,18, used with permission.)

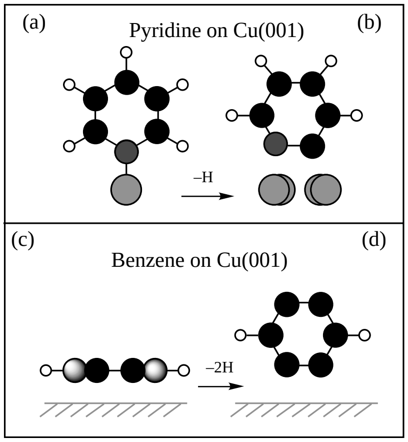


Figure 12. Influence of the sample-substrate interaction on sample orientation. (a) An individual pyridine molecule on a copper (001)-orientation, (Cu(001)) substrate is perpendicular to the surface due to the strong nitrogen–copper atom interaction. (b) An individual pyridine molecule from which a hydrogen atom has dissociated is also perpendicular to a Cu(001) surface but has a shifted location. (c) An individual benzene molecule on a Cu(001) substrate is parallel to the surface but (d) may become perpendicular if hydrogen dissociation occurs. (Adapted from Ref. 16, used with permission.)

Biomolecules, with only nanometer dimensions, always should be deposited on atomically flat single-crystal substrates. Substrates can also be selected to supply electrons to the biomolecule, for positive bias scanning, or to manipulate the biomolecule into a desired position. Another important sample preparation issue is that biomolecules often have multiple available conformations, including globular conformations that self-protect the molecule under nonphysiological conditions. While STM imaging may be performed in vacuum, air, and even in a liquid-filled compartment (liquid cell), the best resolution may be achieved in vacuum, which is a nonphysiological condition. The less physiological the imaging conditions, the more it will be necessary to use special molecular stretching techniques to investigate an open conformation. A special pressure jet injection technique was developed by the Osaka University group to deposit stretched DNA and RNA on single-crystal copper for vacuum STM imaging, without giving them the chance to close into globular conformations (13,14).

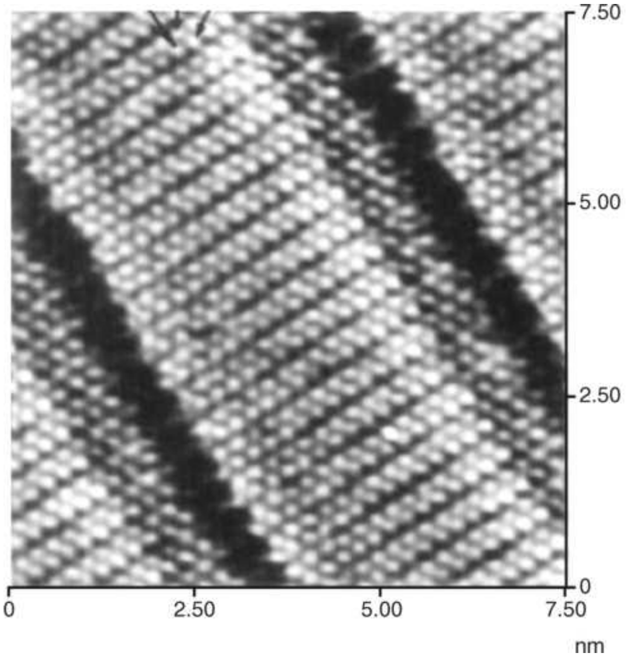


Figure 13. High resolution STM images of an alkane (pentatracontane) monolayer on graphite. (Reproduced from Ref. 19, used with permission.)

IMAGING ARTIFACTS AND DATA RESTORATION USING DECONVOLUTION

Examination of Fig. 5a shows the ring head of retinoic acid as a large blurred bright spot. Greater resolution of detail would clearly be desirable. As in all SPM imaging systems, tip artifacts versus the surface features will limit the resolution of the experiments performed. This is often cited as an ultimate barrier in STM studies of macromolecular structures and in scanning probe microscopy in general (27). It is therefore necessary to develop techniques for deconvolution of STM tip artifacts for enhancing the resolution of measured STM image data.

A commonly used approach for data restoration or eliminating the smearing effect of tip sample interaction is to assume that the observed signal is a convolution of the true image and the probe response function (PRF). The following equation gives a general degradation model due to the convolution of tip artifacts with true data resulting in the measurement *g(x,y)*. Neglecting the presence of the additive noise, the data can be modeled as

$$g(x,y) = f(x,y) * h(x,y) = \sum_{n,m} f(n,m)h(x-n,y-m)$$

where *g(x, y)*, *f(x, y)*, and *h(x,y)* are the observed or raw signal, true image, and PRF, respectively. One can then use deconvolution methods to extract the true image from the knowledge of measured data and probe PRF.

Theoretically, the probe response function is derived from the underlying physics of the tip sample interaction process. Hence, there is a need for a theoretical model for the tip sample interaction. Recent advances in formulation and modeling of tip sample interactions allow development of accurate compensation algorithms for deconvolving the effect of tip-induced artifacts.

Figure 14 shows an example of applying a deconvolution algorithm on synthetic degraded images. The degraded image in Fig. 14c is generated from a synthetic image in Fig. 14a blurred by a Gaussian PRF in Fig. 14b. Figure 14d shows the enhanced result obtained using deconvolution. Although the theoretical treatment of STM and related SPM techniques provide major challenges because the atomic structures of the tip and sample have to be modeled appropriately, its potential is clear and this is a strongly developing research area at the present time.

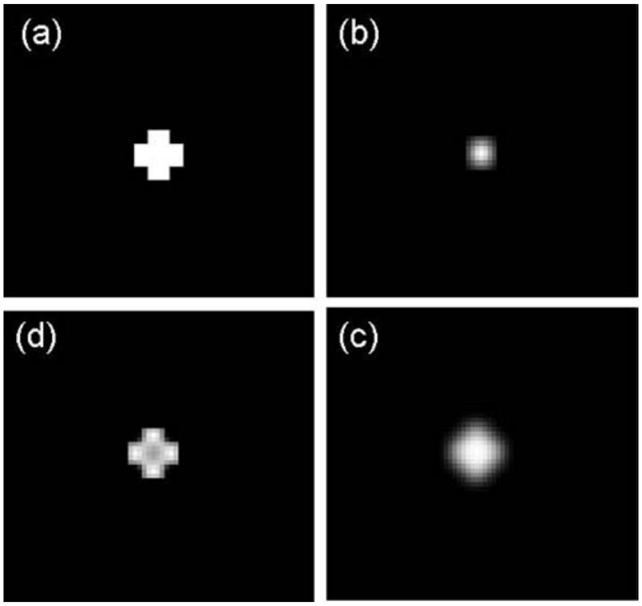


Figure 14. Clockwise from upper left: (a) synthetic true image (b) Gaussian PRF, (c) degraded measurement, and (d) restored image.

CONCLUSIONS

The STM imaging has the highest resolution of all SPM imaging techniques. As such, it would be highly desirable to apply STM to investigations of molecular biology and medicine. An often described difficulty when considering the application of STM to molecular biology is that biological samples are nonconductive. It would be more accurate to describe biological samples as having both local and varying conductivities. Design of STM experiments in which ring systems are exploited, and/or imaging parameters are set for the least conductive portion of the biomolecules may help produce successful imaging results. New research in applications of powerful deconvolution techniques to STM imaging will also open up the field of direct STM investigations of the structure and function of important biomolecules.

BIBLIOGRAPHY

1. Binning G, Rohrer H. Helv Phys Acta 1982;55:726–735.
2. Hansma HG, Oroudjev E, Baudrey S, Jaeger L. TectoRNA and kissing loops: Atomic force microscopy of RNA structures. J Microsc 2003;212:273–279; Sitko JC, Mateescu EM, Hansma HG. Sequence-dependent DNA condensation and the electrostatic zipper. Biophys J 2003;84:419–431; Hansma HG, Vesenka J, Siegerist C, Kelderman G, Morrett H, Sinsherimer RL, Bustanmante C, Elings V, Hansma PK. Reproducible imaging and dissection of plasmid DNA under liquid with the atomic force microscope. Science 1992;256:1180.
3. Hartmann U. Magnetic force microscopy. Annu Rev Mater Sci 1999;29:53–87.
4. Paesler MA, Moyer PJ. Near-Field Optics: Theory, Instrumentation, and Applications. New York: Wiley-Interscience; 1996.
5. Majumdar A. Scanning thermal microscopy. Annu Rev Mater Sci 1999;29:505–585.
6. Beebe TP, Jr., et al. Science 1989;243:370–371.
7. Lee G, Arscott PG, Bloomfiled VA, Evans DF. Science 1989;244:475–478.
8. Driscoll RJ, Youngquist MG, Baldescwieler JD. Atomic scale imaging of DNA using scanning tunnelling microscopy. Nature (London) 1990;346:294–296.
9. Tanaka H, Kawai T. Visualization of detailed structures within DNA. Surface Sci Lett 2003;539:L531–L536.
10. Nishimura M, Tanaka H, Kawai T. High resolution scanning tunneling microscopy imaging of Escherichia coli lysine transfer ribonucleic acid. J Vac Sci Technol B 2003;21:1265–1267.
11. Wikipedia, The free encyclopedia. Available at http://en.wikipedia.org/wiki/RNA.

12. Giege R, Puglisi JD, Floentz C. In: Cohn WE, Moldave K, eds., Progressin Nucleic Acid Research and Molecular Biology, Vol. 45. Amsterdam: Elsevier; 1993.
13. Hadi Zareie M, Lukins PB. Atomic-resolution STM structure of DNA and localization of the retinoic acid binding site. Biochem Biophy Res Commun 2003;303:153–159.
14. Otero R, et al. Proceedings of the 5th Trends In Nanotechnology (TNT04) CMP Cientifica; 2004. Available at http://www.phantomsnet.net/files/abstracts/TNT2004/AbstractKeynoteBesenbacherF.pdf.
15. Available at http://biochem.otago.ac.nz/staff/sowerby/periodicmonolayers.htm. Multiple references listed.
16. Lauhon LJ, Ho W. Single molecule chemistry and vibrational spectroscopy: Pyridine and benzene on Cu(001). J Phys Chem A 2000;104:2463–2467.
17. Image credit in Fig. 11(a): U.S. Department of Energy Human Genome Program, Available at http://www.ornl.gov/hgmis. This image originally appeared in the 1992 U.S. DOE Primer on Molecular Genetics.
18. Image credit in Fig. 11(b): Mathematical Association of America, Available at http://www.maa.org/devlin/devlin0403.html.
19. Claypool CL, et al. Source of image contrast in STM images of functionalized alkanes on graphite: A systematic functional group approach. J Phys Chem B 1997;101:5978–5995.
20. Faglioni F, Claypool CL, Lewis NS, Goddard WA III. Theoretical description of the STM images of alkanes and substituted alkanes adsorbed on graphite. J Phys Chem B 1997;101:5996–6020.
21. Claypool CL, et al. Effects of molecular geometry on the STM image contrast of methyl- and bromo-substituted alkanes and alkanols on graphite. J Phys Chem B 1999;103:9690–9699.
22. Claypool CL, Faglioni F, Goddard WA III, Lewis NS. Tunneling mechanism implications from an STM study of $H_3C(CH_2)15HC=C=CH(CH_2)15CH_3$ on graphite and $C_{14}H_{29}OH$ on MoS_2. J Phys Chem B 1999;103:7077–7080.
23. Villarubia JS. Algorithms for scanned probe microscope image simulation, surface reconstruction and tip estimation. J Res Nat Inst Standards Technol 1997;102: 425–454.

See also BIOSURFACE ENGINEERING; MICROSCOPY, ELECTRON.

MICROSENSORS FOR BIOMEDICAL APPLICATIONS. See CAPACITIVE MICROSENSORS FOR BIOMEDICAL APPLICATIONS.

MICROSURGERY

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

INTRODUCTION

Microsurgery is a specialized surgical technique whereby a microscope is used to operate on tiny structures within the body. Fine precision microinstruments are used to manipulate tissue with or without robotic or computer control. This technique has allowed for significant advances in surgery especially for operations involving the inner ear, eye, brain, nerves, and small blood vessels.

HISTORY OF THE SURGICAL MICROSCOPE

The compound microscope is generally agreed upon to have been invented in the 1590s by the two Dutch opticians Hans and Zacharias Janssen. Their device consisted of a sliding tube with two aligned lenses. In 1624 Galileo Galilei, the famous astronomer and mathematician, demonstrated an inverted telescope to his colleagues of the Lincean Academy. One of them, Giovanni Faber, named it a microscope from the Greek words micro, meaning small, and scope, meaning to aim or shoot. Microscope lenses were improved in the seventeenth century by Antonie van Leeuwenhoek, a Dutch linen draper who originally was interested in counting the number of threads per square inch in his reams of cloth. How he constructed his spherical lenses still remains a mystery to this day. In the eighteen century, fine and course adjustments as well as tube inclination were added by Robert Hooke, who first discovered the cell. The microscope was further improved by Joseph Jackson Lister a British wine merchant, school visitor, and histologist, the father of Lord Joseph Lister whom is credited as stating the era of modern surgery. His innovations included the developed achromatic objective lens corrected for chromatic and spherical aberrations and stands designed to reduce vibrations. His jointly published work with Dr. Joseph Hodgkins in 1827, redefined the understanding at the time of arteries, muscles, nerves, and the brain.

In 1848 Carl Zeiss, a German machinist opened a microscope workshop. Ernst Abbé, a physicist working with Zeiss, derived new mathematical formulas and theories that allowed the optical properties to be mathematically predicted for the first time. Prior lenses had always been made by craftsmen who learned their trade by trial and error. Abbé's advancements enabled Zeiss to become the first mass producer of high quality microscopes.

USE IN SURGERY

Edwin Theodor Saemisch, a German ophthamologist, used loupes in surgery in 1876, but although the microscope was being used in the laboratory medical research environment it was not used the operating room. Zeiss manufactured a binocular microscope specifically designed for dissecting which was used for ophthalmological examinations of the cornea and anterior chamber of the eye. It was not until 1921 that Carl Olof Nylen, a Swedish, otologist and tennis olympian, used his homebuilt monocular microscope for the first time in ear surgery on a case of chronic otis media, a type of ear infection. His monocular microscope was quickly replaced in 1922 by a binocular microscope developed by adding a light source to a Zeiss dissecting microscope by his chief surgeon, Gunnar Holmgren. He used it to treat diseases otosclerosis, the abnormal growth of temporal bone in the middle ear.

Despite these early successes the surgical microscope was seldom used due to its limited field of vision, very short focal distance, poor light quality, and instability. It was not until the 1950s that the surgical microscope started to become more widely adopted. In 1953, Zeis released the Zeiss OpMi 1(Zeiss Operating Microscope Number One), which was specially designed for otological procedures. Its superior coaxial lighting, stability, and ease of operation enabled the advent of tympanoplasty operations to repair ruptured ear drums as well as widespread use in temporal bone surgery.

The success the microscope was having in otology soon spread to other disciplines as well. In the early 1950s, José Ignacio Barraquer adapted a slip lamp to the Zeiss otological surgical microscope for ocular microsurgery. By the 1960s, J. I. Barraquer, Joaquín Barraquer, and Hans Littman of Zeiss, had further modified the surgical microscope and refined microsurgical techniques to make ocular maneuvers in glaucoma microsurgery easier to perform. During this same time frame, Richard Troutman also had Zeiss make a special microscope for his ophthalmic procedures. He made many advances and is credited as adding electric and hydraulic control to surgical microscopes, but possibly his greatest innovation was the first variable magnification, or zoom, surgical microscope.

Around this time neurosurgeons also began using the surgical microscope in the operating room. In 1957, Theodore Kurze removed a neuriloma tumor from the seventh cranial nerve, and then later anastomized it to the hypoglossal nerve. He also developed techniques to use the surgical microscope for aneurysm surgeries. Recognizing that sterilization was a major problem, he developed the use of ethylene oxide gas to sterilize his surgical microscopes for use in the operating room. Dr. R. M. Peardon Donaghy established the first microsurgical training lab at the University of Vermont, where many surgeons were trained. He collaborated with the vascular surgeon Julius Jacobson to remove occlusions from cerebral arteries. Jacobson and his colleague Ernesto Suarez, were responsible for developing small vessel anastomoses techniques. Their procedures required another surgeon's assistance. To meet this need Jacobson invented the diploscope that allowed two surgeons to view the same operative field. Later he and his colleagues worked with Hans Littman of Zeiss to develop a commercial surgical microscope with a beamsplitter enabling two surgeons to operate at the same time. A modern day version of the Zeiss dual head microscope is shown in Fig. 1.

Inspired by the work of the neuroscientists plastic surgeon also started using the surgical microscope. Harold Buncke was one of the first plastic surgeons to use the microscope for digit/limb replantation and free-flap autoplantation. Buncke also developed many of the tools microsurgeons use by borrowing technology from the jewelry, watchmaking, and microassembly industries (Fig. 2.)

The 1960s saw the techniques applied to neurosurgery. Breakthroughs in microneurosurgery included the repair of peripheral nerve injuries, intracranial aneurysm surgeries, embolectomies of middle cerebral arteries, middle cerebral artery bypasses. One on the visionaries of this time was M. G. Yasargil a Turkish neurosurgeon from Switzerland. Trained in Donaghy's training lab he further refined and improved the surgical techniques.

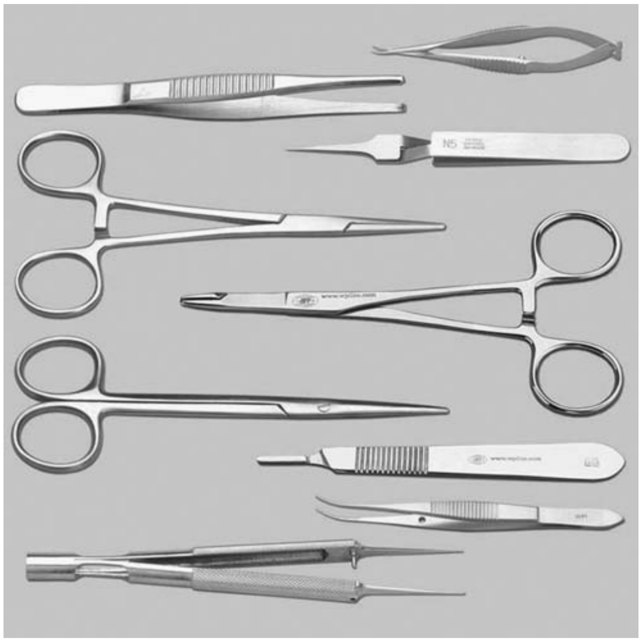


Figure 1. Zeiss OpMi Vario S8 surgical microscope. (Courtesy of Carl Zeiss.)

The next advancements came with the development of minimally invasive surgical techniques. Traditional surgical techniques used relatively large incisions to allow the surgeon full access to the surgical area. This type of operation, called open surgery, enables the surgeon's hands and instruments to come into direct contact with organs and tissue, allowing them to be manipulated freely. These operations are classified as first generation surgical techniques and most surgeons are trained in this manner. While the large incision gives the surgeon a wide range of motion to do very fine controlled procedures, it also causes substantial trauma to the patient. In fact, the majority of trauma is caused by the incisions the surgeon uses to get access to the surgical site instead of the actual

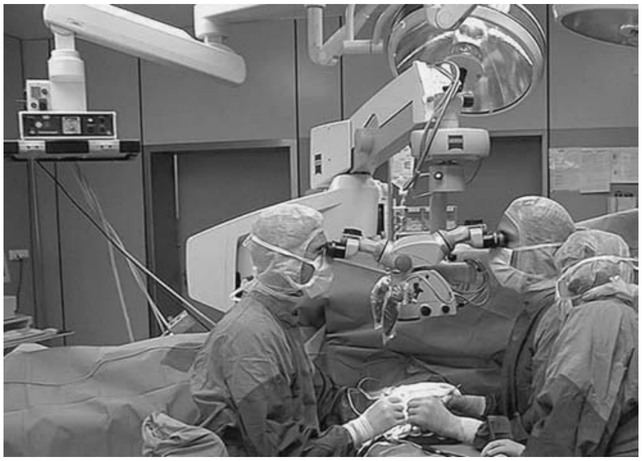


Figure 2. Modern day microsurgical tools. (Courtesy of WPI Inc.)

surgical procedure itself. For example, in a conventional open-heart cardiac operation, the rib cage must be cracked and split exposing the heart muscle. This trauma not only increases pain to the patient, but adds to recovery times increasing hospital stays, in turn increasing costs.

In 1985, Muhe performed the first laparoscopic cholecystectomy, or gallbladder removal surgery with a fiberoptic scope. The technique he performed is commonly called minimally invasive surgery but also goes by other names, such as micro, keyhole, microscopic, telescopic, less invasive, and minimal access surgery. This microsurgical technique is based on learnings from gynecological pelviscopies and arthroscopic orthopedic operations along with the previous advances made in otology, opthamology, neurosurgery, and reconstructive microsurgeries. It has subsequently been applied to many other surgical areas, such as general surgery, urology, thoracic surgery, plastic surgery, and cardiac surgery. These procedures are classified as second generation surgeries as trauma to the patient is drastically reduced by the reducing or eliminating incisions. The shorter hospital stays and faster recovery times for the patient reduce the cost of a minimally invasive procedure 35% compared to its open surgery counterpart.

In a minimally invasive cardiac operation a few small holes, access points, or ports are punctured into the patient and trocars are inserted. A trocar consists of a guiding cannula or tube with a valve–seal system to allow the body to be inflated with carbon dioxide. This is done so that the surgeon has enough room to manipulate his instruments at the surgical site. An endoscope is inserted into one of the trocar ports to allow the surgeon a view the surgical site. Various other surgical instruments, such as clippers, scissors, graspers, shears, cauterizers, dissectors, and irrigators were miniaturized and mounted on long poles so that they can be inserted and removed from the other trocar ports to allow the surgeon to perform the necessary tasks at hand.

While minimally invasive surgery has many advantages to the patient, such as reduced postoperative pain, shorter hospital stays, quicker recoveries, less scarring, and better cosmetic results, there are a number of new problems for the surgeon. The surgeon's view is now restricted and does not allow him to see the entire surgical area with his eyes. While the operation is being performed he must look at a video image on a monitor rather than at his hands. This is not very intuitive and disrupts the natural hand–eye coordination we all have been accustomed to since childhood. The video image on the monitor is also only two dimensional (2D) and results in a loss of our binocular vision eliminating the surgeon's depth perception. While performing the procedure the surgeon does not have direct control of his own field of view. A surgical assistant holds and maneuvers the endoscopic camera. The surgeon has to develop his own language to command the assistant to position the scope appropriately, which often leads to orientation errors and unstable camera handling, especially during prolonged procedures. Since the images from the camera are magnified, small motions, such as the tremor in a surgical assistant's hand or even their heartbeat can cause the surgical team to experience motion induced nausea. To combat the endoscopic problems, some surgeons choose to manipulate the endoscope themselves. This restricts them to using only one hand for delicate surgical procedures and makes procedures even more complicated.

The surgeon also loses the freedom of movement he has in open surgery. The trocar ports are fixed to the patient's body walls by pressure and friction forces. This constrains the instrument's motion in two directions and limits the motion of the tip of the instrument to four degrees of freedom (in–out, left–right, up–down, and rotation). The trocars also act as pivot points and cause the surgical instruments to move in the opposite direction to the surgeon's hands. When the surgeon is moving left, the image on the monitor is moving to the right. The amount of this opposite movement also depends on the depth of the introduction of the instrument. Again because of the pivot point the deeper an instrument is inserted into the body the more the surgeon's movement is amplified. Even a small movement made by the surgeon on the outside of the patient can translate to a very large movement on the inside of the patient. The seals and valves in the trocars also impede movements which hinders the smoothness of motions into and out of the patient and greatly reduces the already limited tactile feedback the surgeon experiences. These movement behaviors and lack of tactile feedback are counter to what the surgeon is used to in open surgery and require long training to develop the technical skills to perform these operations.

Performing a minimally invasive procedure has been likened to writing your name holding the end of an 18 in. (45.72 cm) pencil (1). The surgeon loses three-dimensional (3D) vision, dexterity, and the sense of touch. The instruments are awkward, counterintuitive, and restricted in movement. The lack of tactile feedback prevents the surgeon from knowing how hard he or she is pulling, cutting, twisting, suturing, and so on. These factors cause a number of adjustments to be made by the surgeon, which requires significant retraining on how to do the procedures in a minimally invasive manner. These difficulties encountered by the surgeon cause degradation in surgical performance compared to open surgery which limits surgeons to performing only simpler surgical procedures.

In an attempt to address some of these shortcomings and allow the surgeon more control during operations a third generation of surgical procedures, robotic surgery was developed. Although these types of procedures are commonly referred to as robotic surgery, the operations themselves are not completely automated and are still carried out by a surgeon. For this reason, robotic surgery is also referred to as computer aided or computer assisted surgery.

The robotic technology was originally developed for telerobotic applications in the late 1980s for the Defense Advanced Research Project Administration (DARPA) by researchers at SRI International. The surgeon of the future would allow surgeons from remote command centers to operate on injured soldiers in the battlefield. In 1995, this technology was spun off into a new company named Intuitive Surgical to commercialize the technology for use in the hospital environment. Near the same time Dr. Yulan Wang was developing robotic technology for NASA to allow

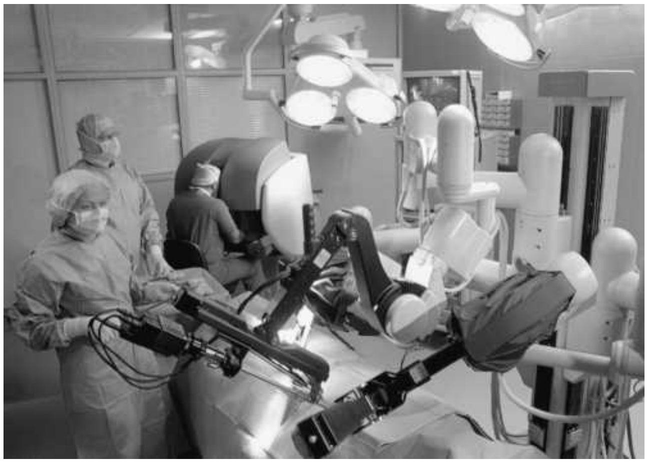


Figure 3. Intuitive Surgical da Vinci robotic surgery system. (Copyright ©2005 Intuitive Surgical, Inc.)

surgeons on earth to deal with medical emergencies on the international space station. He formed Computer Motion in 1989. Both of these companies merged in 2003, and Intuitive Surgical is now the leader in Robotic Surgery. In 2002, Dr. Fredric Mol, one of the original founders of Intuitive Surgical, founded Hansen Medical which brings computerized robotic control of catheters to electrophysiology and interventional cardiac procedures.

Current robotic surgery systems have a number of benefits over conventional minimally invasive surgery. Figure 3 shows an Intuitive Surgical da Vinci robotic system. In this arrangement the surgeon sits comfortably at a computer console instead of having to stand throughout the entire procedure, which can last up to 5 h long. A three-armed robot takes his place over the patient. One arm holds an endoscope while the other two hold a variety of surgical instruments. The surgical team can also look at a video monitor to see what the surgeon is seeing. The surgeon looks into a stereo display in much the same way as looking though a surgical microscope and manipulates joystick actuators located below the display. This simulates the natural hand–eye alignment he is used to in open surgery, (Fig. 4). Since computers are used to control the robot and are already in the operating room, they can be used to give the surgeon superhuman like abilities. Accuracy is improved by employing tremor cancellation algorithms to filter the surgeon's hand movements. This type of system can eliminate or reduce the inherent jitter in a surgeon's hands for operations where very fine precise control is needed. Motion scaling also improves accuracy by translating large, natural movements into extremely precise, micromovements. A wide variety of surgical instruments or end effectors are available including graspers, cutters, cauterizers, staplers, and so on. Both companies provide end effectors that have special wrist like joints at their tips enabling full seven degree of freedom movements inside the patient, (Fig. 5), but still lack tactile feedback.

These robotic advances allow surgeons to perform more complex procedures, such as reconstructive cardiac operations like coronary bypass and mitral valve repair that cannot be performed using other minimally invasive techniques.

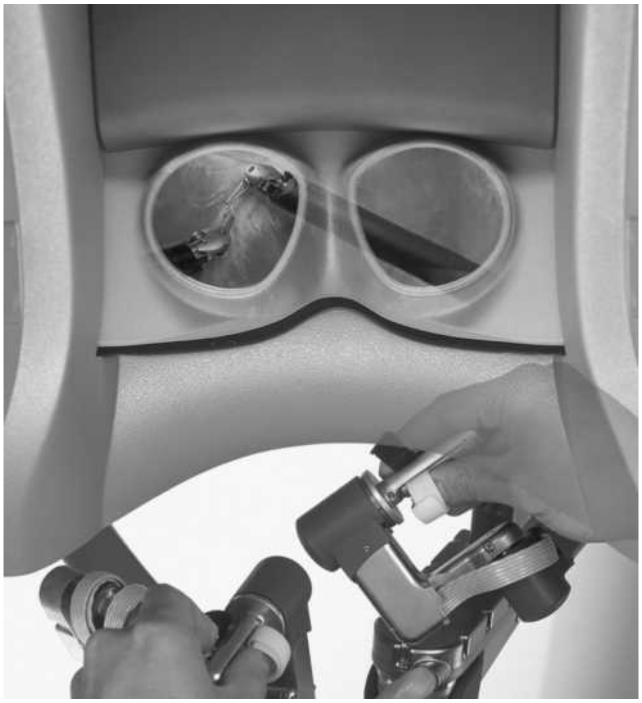


Figure 4. Intuitive Surgical stereo display and joysticks. (Copyright ©2005 Intuitive Surgical, Inc.)

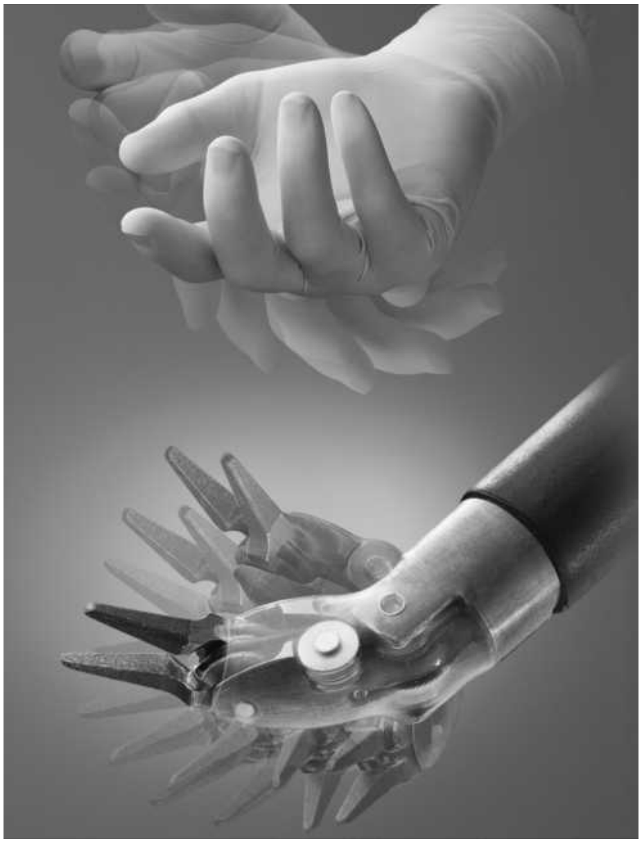


Figure 5. Multidegrees-of-freedom end effector. (Copyright ©2005 Intuitive Surgical, Inc.)

MEMS

Around the same time that minimally invasive surgery was being developed, there was a turning point in microelectromechanical systems (MEMS). This a technology was developed from the integrated circuit industry to create miniature sensors and actuators. Originally, these semiconductor processes and materials were used to build electrical and mechanical systems, but have now expanded to include biological, optical, fluidic, magnetic, and other systems as well. The term MEMS originated in the United States and typically contain a moving or deformable object. In Europe, this technology goes by the name microsystems technology or microstructures technology (MST) and also encompasses the method of making these devices, which is referred to as micromachining. In Japan and Asia MEMS are called micromachines when mechanisms and motion are involved. The MEMS devices first were used in medical applications in the early 1970s with the advent of the silicon micromachined disposable blood pressure sensor (2), but it was not until the mid-1980s when more complicated mechanical structures, such as gears and motors, were able to be fabricated.

FABRICATION TECHNOLOGIES

The fabrication of MEMS devices is based on the merger of semiconductor microfabrication processes and micromachining techniques to create the desired microstructural components. There are four major processes that are used to fabricate MEMS devices: bulk micromachining, surface micromachining, LIGA, and precision machining. Combinations of these technologies are what allow MEMS to be highly miniaturized and integratable with microelectronics. These processes are very sensitive to impurities and environmental conditions such as temperature, humidity, and air quality. Typically, these fabrication steps are performed inside a cleanroom (Fig. 6). Bulk micromachining, surface micromachining, and LIGA have the added advantage of being able to be batch fabricated. This allows many devices to be made in parallel at the same time greatly reducing device cost.

Bulk micromachining utilizes wet- or dry-etch processes to form 3D structures out of the substrate. These subtractive processes produce isotropic or anisotropic etch profiles in material substrates, which are typically but not limited to silicon wafers. Bulk micromachining can create large MEMS structures on the micrometers (μm) to millimeters (mm) scale (tens of μm-to-mm thick). Commercial applications of bulk micromachining have been available since the 1970s. These applications include pressure sensors, inertial sensors such as accelerometers and gyros, and microfluidic channels and needles for drug delivery.

In surface micromachining, MEMS are formed on the surface of the substrate using alternating layers of structural and sacrificial materials. These film materials are repeatedly deposited, patterned, and etched to form structures that can then be released by removing sacrificial layers. The release process allows for the fabrication of complex movable structures that are already assembled, such as motors, switches, resonators, cantilevers, and so on. Surface micromachined structures are typically limited to thicknesses of 2–6 μm and because they use much of the same technology as is used in the integrated circuit industry are readily integrated with electronics. Because so much technology is shared with the IC industry silicon wafers are typical substrates with thousands of devices being able to be fabricated at once (Fig. 7).

Figure 6. Cleanroom fabrication facility. (Courtesy of Intel Corp.)

Lithographie, Galvanik, Abformung (LIGA) is a German acronym that means lithography, electroforming, and molding. The technology was originally developed in the late-1970s to fabricate separation nozzles for uranium enrichment. This technology uses X rays to fabricate devices with very high aspect ratios. A synchrotron radiation source is used to define small critical dimensions in a poly(methyl methyacrylate) (PMMA), mold that can then be electroplated to form high aspect ratio metallic structures. Many parts can be batch fabricated in this manner, but assembly is usually still a serial process.

Precision machining technology, such as micro-EDM (microelectro discharge machining), laser micromachining, and micro stereo lithography, is also used to form complex structures out of metal, plastic, and ceramics that the previous fabrication technologies may be incapable of. Precision machining is typically a serial process, but is

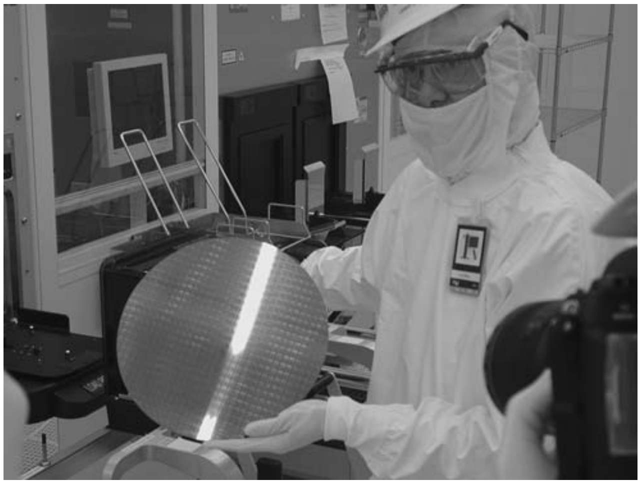


Figure 7. Silicon wafer in the fabrication process. (Courtesy of Intel Corp.)

often better able to deal with the varied shapes and substrates of microsurgical instruments.

Micro EDM is a form of spark machining used to shape conductive materials, such as silicon and metals. An EDM erodes material by creating a controlled electric discharge between an electrode and the substrate. It is a noncontact process and there is no direct mechanical cutting force applied to the substrate. Dielectric fluid is used to remove the erosion particles, as well as to keep the substrate material from oxidizing. Micro-EDMs can be used to make holes, channels gears, shafts, molds, dies, stents, as well as more complex 3D parts such as accelerometers, motors, and propellers (3).

Lasers can be used to both deposit and remove material. Laser ablation vaporizes material through the thermal noncontact interaction of a laser beam with the substrate. It allows for the micromachining of silicon and metals, as well as materials that are difficult to machine using other techniques such as diamond, glass, soft polymers, and ceramics. Laser direct writing and sintering is a maskless process where a laser beam is used to directly transfer metal materials onto a substrate. This can be used to form metal traces on nonplaner surfaces, which reduces the need for wires on surgical tools (4).

Micro stereo lithography processes generate 3D structures made out of ultraviolet (UV) cured polymers. It is an additive process where complex 3D structures are made from stacks of thin 2D polymer slices that have been hardened from a liquid bath. Conventional systems were limited in that they were a serial process where only one part could be made at a time. MicroTEC has developed a batch fabricated wafer level process called rapid material product development (RMPD), which is capable of constructing structures out of 100 different materials including plastics, sol–gels, and ceramics (5).

APPLICATIONS

The inclusion of MEMS technology in microsurgery, will allow for smaller more miniaturized surgical tools that not only overcome many of the limitations of microsurgical procedures, but allow for new more advanced operations to be performed. MEMS are just now being incorporated into microsurgical tools and coming on the market. Most are still at the research level, but the industry is moving in this direction as the need for smaller smarter tools increases.

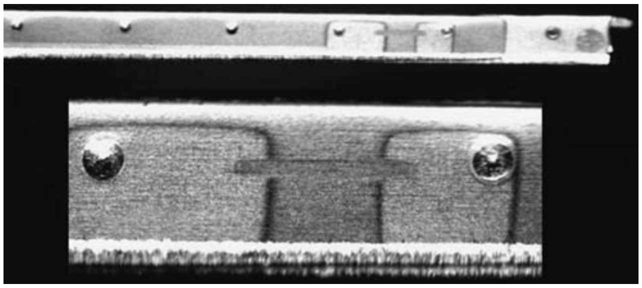


Figure 8. Strain gauges fabricated on surgical sharps. (Courtesy of Verimetra, Inc.)

HAPTIC FEEDBACK

One of the key areas for improvement in microsurgery is tactile feedback. The lack of tactile sensing limits the effectiveness these procedures. Recent work in robotic feedback for minimally invasive surgery has concentrated on force feedback techniques using motors and position encoders to provide tactile clues to the surgeon. In these approaches, the sense element is far removed from the sense area. Verimetra, Inc. has developed strain gauge force sensor fabrication technology which uses the surgical tools themselves as a substrate (6). Prior efforts have focused on fabrication of sensors on silicon, polyimide, or some other substrate followed by subsequent attachment onto a surgical tool with epoxy, tape, or some other glue layer. Attaching a sensor in this manner limits performance, introduces sources of error, limits the sensor's size, and further constrains where the sensor can be placed. By eliminating glue and adhesion layers improved sensitivity and reduces errors due to creep. Figure 8 shows strain gauges fabricated on surgical sharps. Figure 9 is a cut away SEM image of a strain gauge and temperature sensor embedded inside of a robotic microforcep. While this microfabrication technology is an improvement in sensor technology, wires are still used to connect the sensor to the outside world. Reliability and the added complexity of adding wires to surgical end effectors with high degrees

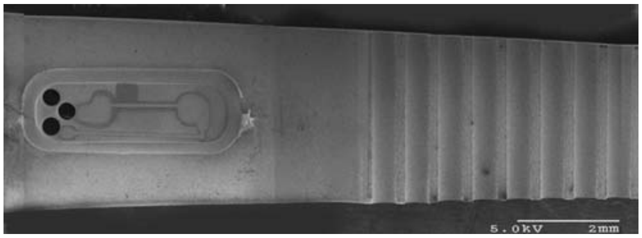


Figure 9. Strain gauges and temperature sensor embedded in robotic microgripper. (Courtesy of Verimetra, Inc.)

of freedom limit the effectiveness of the technology. Short-range wireless technology compatible with the operating room environment need to be developed to overcome these limitations.

Recently tactile feedback has been shown to be able to be added to noncontact lasers. Based on optical distance measurements, the systems synthesize haptic feedback through a robotic arm held by the surgeon when the focal point of the laser is coincident with a real surface. This gives the operator the impression of touching something solid. By increasing the power of the laser such a system could also be used for cutting or ablation.

TISSUE SENSING

Taking haptic feedback one step further is the ability to distinguish between different types of tissue in the body. Tissue sensing is of vital importance to a surgeon. Before making an incision into tissue, the surgeon must identify what type of tissue is being incised, such as fatty, muscular, vascular, or nerve tissue. This becomes more complicated because the composition and thickness of different human tissues varies from patient to patient. Failure to properly classify tissue can have severe consequences. For example, if a surgeon fails to properly classify a nerve and cuts it, then the patient can suffer effects ranging from a loss of feeling to loss of motor control. If a neurosurgeon cuts into a blood vessel while extracting a tumor severe brain damage may occur. The identification and classification of different types of tissue during surgery, and more importantly during the actual cutting operation, will lead to the creation of smart surgical tools. If a surgical tool senses that it is too close to or about to cut the wrong type of tissue it can simply turn itself off.

Verimetra, Inc. has developed a device called the data knife, Fig. 10. It is a scalpel, which is outfitted with different strain sensors along the edges of the blade to sense the amount of force being applied. The resistance of the tissue is one of the signals used for classifying tissue. Pressure sensors are used to measure the characteristics of material surrounding the blade. The pressure of the surrounding fluid can be used to help classify the type or location of tissue. Electrodes are used to measure the impedance of different types of tissue, as well as being used to identify nerves by picking up their electrical signals. The tool provides the real-time feedback surgeons have been asking for during surgery, and can also be used to record data for later use for tracking purposes.

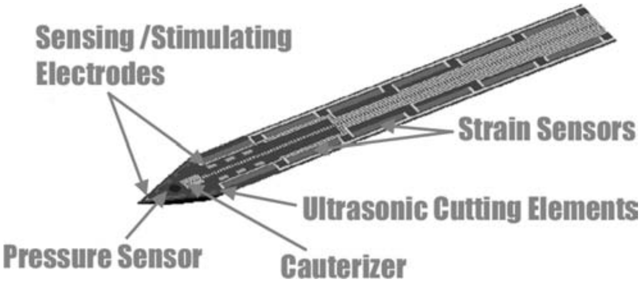


Figure 10. Data Knife smart scalpel. (Courtesy of Verimetra, Inc.)

Sensing the density of tissue can also be used to assist the surgeon in identifying tissue. In open cardiac bypass operations, the surgeons insert their hands inside the body to palpate arteries. For cardiac bypass surgery, surgeons select the bypass location by feeling where the fat and fatty plaque is located in your arteries with their fingers. The lack of tactile feedback in minimally invasive surgery, prevents them from using this technique. The MEMS devices have been developed for the palpation of tissue using strain gauges (7), diaphragms (8), micropositioners (9,10), and load cells (11) and have shown the ability to measure blood pressure, pulse, different kinds of arterial plaque, and distinguish between colon, bowel, stomach, lung, spleen, and liver tissue.

Piezoelectric transducers can also be used to measure density. Macroscale transducers are frequently used in imaging applications to differentiate between tumors, blood vessels, and different types of tissue. These transducers both emit and receive sound waves. By vibrating at a high frequency sound waves are emitted in the direction of the object of interest. The density of the impinged object can then be measured based on the signal that is reflected back by that object. Sound waves are reflected off the interfaces between different types of tissue and returned to the transducer. Present ultrasound systems tend to be large and are not well suited for incorporation into microsurgical devices. The MEMS technology is well suited for this application and many ultrasonic MEMS sensors have been developed for imaging (12–16).

Microelectromechanical systems ultrasound devices for density measurements are shown in Fig. 11. They have been shown to be able to detect the location of bone in tissue and are being applied to atrial fibrillation surgeries. Atrial fibrillation is what causes irregular heartbeats and leads to one out of every six strokes. Drugs can be used to treat this condition, but have dangerous side effects including causing a switch from atrial fibrillation to the more dangerous ventricle fibrillation. Pacemakers and other electrical control devices can be used, but they do not always work for all patients. The most effective treatment is the surgical

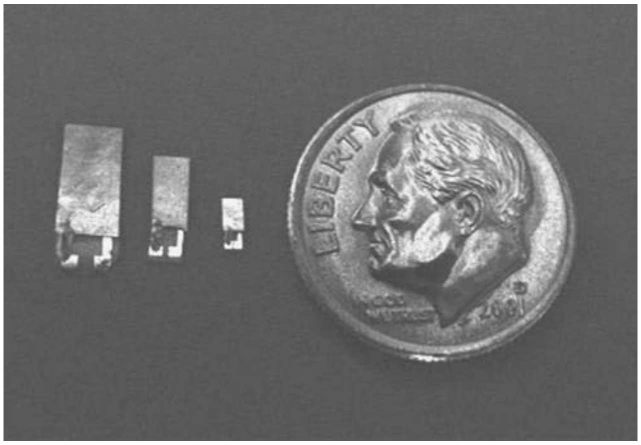


Figure 11. Ultrasound transducers next to a dime. (Courtesy of Verimetra, Inc.)

MAZE procedure, but it is an incredibly invasive treatment. The patient is put on a heart lung machine and then their heart is stopped. Next, the surgeon takes a scalpel and actually cuts all the way through the heart making lesions, which physically separate the heart muscle. These lesions break the electrical connections in the heart. The heart is then sewn back together. Recently, there have been a variety of different methods used to address the problem. Instead of physically cutting all the way through the heart with a scalpel, surgeons are using radio frequency, microwave, cryo, and laser energy to create transmural lesions. Transmurality means that the lesions go all the way through the tissue, breaking the heart's electrical connections. One of the problems surgeons encounter is to know how deep the ablation is or if it is transmural. If the lesions are not completely transmural or just partially transmural then the undesirable electrical signals may still be transmitted to the heart muscle. The MEMS ultrasound technology or micromachined electrodes can be used to measure the transmurality.

Temperature can be used to detect if a surgical device is close to a blood vessel, or if the surgical tool is in a diseased or infected area. Temperature can also be used to monitor the usage of a surgical device, by monitoring the time at which the device is at body temperature. Usage is just one of many areas where Auto-ID technologies will benefit microsurgery (17). They can be used to make sure that only the correct surgical tool is used for a procedure and if that tool has been properly sterilized. Keeping track of how many times, how long, and what was done with a surgical tool will improve the reliability and effectiveness of surgical procedures and will greatly impact the entire medical industry.

TRACKING SYSTEMS

Traditionally, a surgeon uses an endoscope in minimally invasive surgery to determine where the surgical instruments are located in the patient. The view the surgeon has of the surgical area is not ideal and the position of surgical instruments outside of the camera view is not known. Ideally, the surgeon would like to know the position and orientation of each of his instruments. Computer-aided surgery has enabled the surgeon to overlay magnetic resonance imaging (MRI) or computed artial tomography (CAT) scan images of the patient with position and orientation data taken during surgery to create 3D models that the surgeon can use to better visualize the surgical procedure. Computers can be used to simulate the procedure beforehand allowing the surgeon to practice difficult operations ahead of time.

Current technology in this area is predominately optical in nature. Markers are placed on the ends of the surgical instruments that are located outside of the body, as well as on specific locations on the patient's body. A computer registers the location of the surgical tools with the reference markers on the patient so that images of the patient's body can be aligned with the surgical tools. This is done through the use of visible and infrared (IR) cameras. The tips of the surgical tools that are located inside of the body are then extrapolated. The markers must not interfere with the surgery in any way, and therefore should be as small and lightweight as possible. While these systems are wireless and do not have cords that can get tangled on the surgeon or on the surgical tools, there must be an unobstructed path from the markers to the camera systems. The surgeon must be careful not to block the markers themself or with other surgical instruments. Precision is compromised because the location of the surgical tips is extrapolated and does not take into account bending of the surgical tools. Markers on the outside of the body do not take into account compression of the tissue.

MEMS based ultrasound tracking systems have been developed to address these issues (18). Constellations of ultrasound sensors can also be placed on the surgical tools themselves to determine position thereby eliminating errors from extrapolation. Reference markers can now be placed inside of the body, closer to the surgical area so that they are less affected by compression and movement of the patient.

Position and orientation can also be estimated using accelerometers and gyroscopes. The signal outputs can be integrated to determine or predict the distance traveled by a surgical tool. Conventional MEMS accelerometers have accuracies in the milligram range, which are not sufficient for measuring accurately the relatively small displacements made during surgery (19). More accurate inertial sensors need to be developed before they can be integrated into surgical tools.

Magnetic field sensors can also be used to determine position and orientation of surgical tools (20,21). A three axis magnetoeffect sensor has been developed for determining the location of catheters. Currently this is done by continually taking X rays of the patient. However X rays only provide a 2D snapshot, a 3D image would be more helpful for navigation.

EYE SURGERY

The leading cause of vision loss in adults > 60 are cataracts. The word cataract comes from the Greek meaning waterfall and was originally thought to be caused by opaque material flowing, like a waterfall, into the eye. The condition is actually caused by the clouding of the eye's intraocular lense. In the eye, light passes through the lens that focuses it onto the retina. The retina converts the light into electrical signals that are then interpreted by the brain to give us our vision. The lens is a hard crystalline material made mostly of water and protein. The protein is aligned in such a way to allow light to pass through and focus on the retina. When proteins in the lens clump together, the lens begins to cloud and block light from being focused on the retina. This causes vision to become dull and blurry, which is commonly referred to as a cataract.

Much like other surgery in other parts of the body, cataract surgery has followed down the minimally invasive path for the same benefits. Cataract surgery is one of the earliest known surgical procedures. The earliest evidence is the written Sanskrit writings of the Hindu surgeon

Susrata dating from the fifth century BC. He practiced a type of cataract surgery known as couching or reclination, in which a sharp instrument was inserted into eye and pushed the clouded lens out of the way. This displacement of the lens enabled the patient to see better. Although vision was still blurred without corrective lenses, many famous people underwent this procedure including the artists Michelangelo, Rembrandt, and Renoir. Couching was still performed until the mid-twentieth century in Africa and Asia.

In 1748, Jacques Daviel of Paris introduced extracapsular surgery where the lens was removed from the eye. Later, very thick pairs of glasses were used to focus the light onto the retina and restore sight, but the glasses were cumbersome and caused excessive magnification and distortion.

By 1949, Dr. Harold Ridley of England, used PMMA as the first intraocular lens. He discovered that PMMA was biocompatible with the eye while treating WWII fighter pilots whose eyes were damaged by shattering plastic from their windshields. In the 1960s and 1970s, extracapsular surgery became the preferred treatment. A large incision (10–12 mm) was made in the eye to remove and replace the lense. This procedure minimized problems with image size, side vision, and depth perception, but the large incisions required longer hospitalization, recovery time, and stitches.

Today, cataracts are removed with a procedure called phacoemulsication with ~1.5 million operations performed yearly. A hollow ultrasonically driven titanium needle is inserted into the anterior chamber of the eye. Ultrasonic energy is then used to liquefy the hard lens and it is then aspirated out of the eye. A foldable lens made of acrylic or silicone is inserted through a (1–3 mm hole) as a replacement. Since the incision size has been reduced compared to conventional extracapsular surgery, hospitalization, general anesthesia, sutures and bandages have all been eliminated. The reduction in incision size has also reduced the risk of infection and postoperative refractions.

During the procedure the surgeon cannot see directly under the needle as the lens is broken up and aspirated. A thin clear membrane or capsule surrounds the lense. The posterior capsule tissue underneath the lens is very delicate and easily cut compared the crystalline lens. To prevent the soft underlying tissue from damage requires a skilled surgeon whom has performed many procedures. If the posterior capsule is ruptured it can lead to glaucoma, infection, or blindness. As the size of the incision has decreased, heat damage to surrounding tissue from the ultrasonic tip has increased that can alter the characteristics of the tissue and change its appearance. In addition positive intraocular pressure must be maintained by balancing the flow of infusion fluid at positive pressure and the aspirated cataract lens fragments. If pressure is not maintained the anterior chamber can collapse. Pressure is currently maintained by sensors located many feet from the surgical area. This distance creates delays in the feedback loop, which can cause dangerous pressure fluctuations leading to damage to the iris and cornea.

Recently, micromachined silicon ultrasonic surgical tools for phacoemulsifiaction have been developed by Lal's research group (22,23) (Fig. 12). Piezoelectric material is attached to a micromachined silicon needle. The needle has a fluid channel for aspiration as well as a horn for amplifying the ultrasonic displacement. These silicon devices are able to generate higher stroke velocities and lower heat generation than their conventional titanium counterparts. High levels of integration has been achieved by integrating pressure and flow sensors directly on the needle for maintaining intraocular pressure, reducing delays and making the phacoemulsification procedure safer.

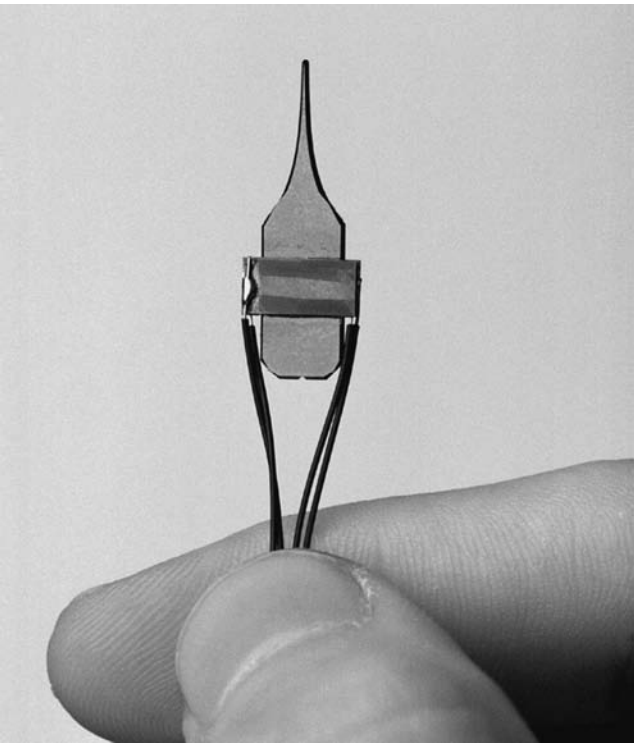


Figure 12. Ultrasonic phacoemulsification tool. (Courtesy of Jeff Miller/University of Wisconsin-Madison.)

To prevent damage to the posterior capsule a piezoelectric sensor has been integrated into a phacoemulsification hand piece and undergone clinical trials (24). The device can determine tissue hardness by measuring the impressed loading on the needle tip or by monitoring the resonant frequency at which the ultrasonic system oscillates. Both of these approaches have proven successful in determining when a hard to soft tissue transition has occurred during a phacoemulsification procedure. This technology enables a surgeon to get real-time feedback on the type of tissue he is cutting, and can be applied to other types of surgical procedures such as tumor extraction as well.

Insertion of a replacement lense requires precise movements by the surgeon. Serious postoperative vision problems may occur if the lens is inserted incorrectly and needs to be removed. Precision piezoelectric micromotors have been developed for intraocular delivery of replacement lenses after cataract removal (25). These inchworm actuators use a glider and clamp arrangement to generate large forces over small displacements. An electrostatic clamp made of an oxide dielectric layer sandwiched between two silicon wafers layer locks the micromotor in place while a PZT actuator generates the force. The inertia of a mass is used to move the clamp. Forces of 3.0 *N* and step sizes of 100 nm to 10 μm have been reported.

In eye surgery there are many times when precision cutting is required. Highly sharpened steel, ceramic, or diamond scalpel blades are used, but are expensive to produce costing up to $1000 a piece. Disposable silicon micromachined scalpels are an attractive alternative. They can be batch fabricated to reduce cost and sharpened to an atomic edge along their crystal planes. They are already being used in Russia at the Fyodorov Eye Care Center in nonpenetrating deep sclerectomy operations for the treatment of glaucoma (26). BD (Bekton, Dikenson, and Company) is producing Atomic Edge silicon blades for cataract surgery (27). Soon they will be used for eye operations and eventually migrate to procedures on other parts of the body. Smaller incisions made by sharper blades result in less bleeding and tearing. An added advantage of silicon blades is that sensors and electronics can be directly fabricated on them during fabrication. Integrating cauterizing electrodes on the blade itself will prevents the patient from bleeding as well as let the surgeon more clearly see the surgical area.

CATHETERS/GUIDEWIRES/STENTS

Cardiac catheterizations can be referred to as a noninvasive surgical procedure. Specialized tubes or catheters are threaded up through blood vessels in the groin, arm, or neck to an area of the body which needs treatment. The problem is then treated from the inside of the blood vessel. The advantage of these approaches is that the procedures require very small incisions, hospital stays are usually one night or less and the discomfort and recovery times afterwards are minimal. For patients with more complicated ailments, catheter treatments can be used in combination with minimally invasive or open surgery to give the best possible results at the lowest possible risk. Catheters, guidewires, and stents currently represent the most widespread use of MEMS technology in surgery.

Diagnostic catheterizations, which are used to measure pressure in different parts of the body, take blood samples, and to perform detailed angiograms of the heart, can be performed by injecting X-ray dye through the catheters. The MEMS pressure sensors are now commonly found on catheter devices and are the most mature MEMS technology in this area. Even smaller designs are being sold for placement on guidewires, such as those made by Silex Microsystems, shown next to a 30 gauge needle (Fig. 13). Each sensor is but 100 μm thick, 150 μm wide, and 1300 μm long (28). MEMS ultrasound sensors are also starting to be used for both forward looking (16) and side looking intravascular imaging (12,13).

To provide the doctor with more information to make better decisions, additional MEMS sensors are needed to gather additional data for the diagnosis, monitoring of procedures, as well as for checking results of completed operations. Many other types of MEMS sensors are being researched to measure blood flows, pressures, temperatures, oxygen content, and chemical concentrations for placement on diagnostic catheters (29,30).

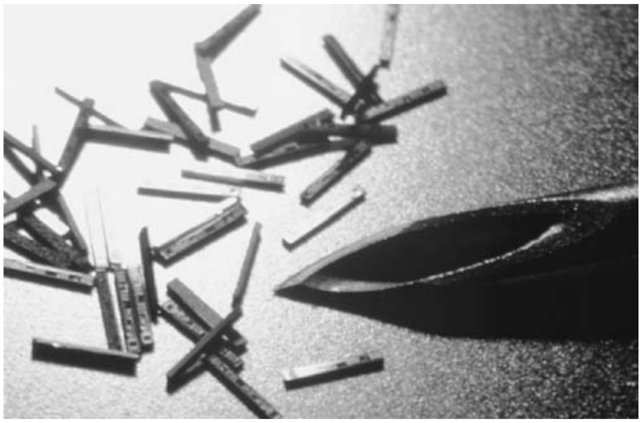


Figure 13. MEMS pressure sensors next to a 30-gauge needle (Courtesy of Silex Microsystems, Jarfalla, Sweden.)

Heart disease continues to be the leading cause of death in the United States. Interventional catheterization is an increasingly more common way to treat blood vessels which have become occluded (blocked) or stenotic (narrowed) by calcified artherosclerotic plaque deposits. Blood vessels that have become occluded or stenotic may interrupt blood flow, which supplies oxygen and cause heart attacks or strokes. Occluded or stenotic blood vessels may be treated with a number of medical procedures including angioplasty and atherectomy. Angioplasty techniques, such as percutaneous transluminal coronary angioplasty (PTCA), also known as balloon angioplasty are relatively noninvasive methods of treating restrictions in blood vessels. A balloon catheter is advanced over a guidewire until the balloon is positioned in the restriction of a diseased blood vessel. The balloon is then inflated compressing the atherosclerotic plaque. Frequently, the wall of the vessel is weakened after inflation and a metal stent in is expanded and deployed against the plaque. The stent helps keep the vessel open. During an atherectomy procedure, the stenotic lesion is mechanically cut or abraded away from the blood vessel wall using an atherectomy catheter.

Microelectrochemical systems pressure sensors can be used to measure the pressure in the balloons during inflation, to make sure damage due to over inflation is minimized. The MEMS temperature sensors can be integrated on catheters and guidewires to determine the location of inflamed plaque. The inflammation causes artery walls in the damaged area to have an increased temperature up to 3°C higher than healthy tissue. Approximately 20–50% of all patients undergoing these therapeutic procedures to clear blocked coronary arteries will suffer restenosis (reblockage) within 6 months of the initial procedure. Drug coated stents have significantly lowered these rates and have been approved for use in Europe for a few years. They are expected to be approved in the United States later this year. The MEMS laser micromachining technology is used in the fabrication of conventional stents and drug coated stents (31). Stainless steel micromachining technology has also been developed at the University of Virginia for piercing structure drug delivery/gene therapy stents for the treatment of restenosis (32). There is potentially a large opportunity for MEMS in embedding sensors into stents to create a smart stents, which would be able to alert doctors when restenosis occurs or other problems occur (33). The MEMS rotary cutting devices have been fabricated for atherectomy procedures (34), but are not widely used because cut up plaque particles can flow downstream and cause strokes.

FETAL SURGERY

Fetal surgical techniques were first pioneered at the University of California San Francisco (UCSF) in the 1980s to operate on babies while they were still in the womb. The success rate of treating certain birth defects is higher the earlier they are addressed in fetal development. Initially open surgical techniques were used with an incision through the mother's abdomen to allow direct access for the surgeon's hands. After the surgery was complete, the womb was sutured and the mother delivered the baby weeks or months later. In 1981, the first fetal urinary tract obstruction procedure was performed at UCSF. Lately, minimally invasive surgical and robotic surgical techniques have been used that have reduced the risk of complications and premature labor. Other fetal procedures have also been developed to treat cojoined twins, hernias, spina bifida, tumors, and heart defects. One area of interest is in fetal heart.

The development of the cardiovascular system in a developing fetus is typically completed by the twelfth week of gestation. At this stage primary heart defects are small and if repaired will prevent defects from becoming more severe as the heart changes to adapt to the normalization blood flows and pressures in the later periods of gestation.

Fetal heart surgery can be used to treat hypoplastic heart syndrome (HHS), which was often considered fatal. This syndrome causes the heart's left side to fail to grow properly and is caused by an obstruction which restricts blood flow. The heart requires the mechanical stress of blood flow to grow properly, and thus fails to develop normally. Today, three open surgeries are required to allow the remaining single (right) ventricle to support both the pulmonary and systemic circulations. The long-term survival for these patients into adulthood is significantly $< 50\%$. Preserving both ventricles would result in the best chances of long-term survival.

An interventional catheter can be used to treat HHS in a minimally invasive manner. The balloon catheter must first penetrate in through the mother's abdominal wall, the placenta and then the fetus's chest into its heart. It must then locate the fetus's tiny heart and expand to open the blockage. Afterward, the surgeon needs to know whether the operation has succeeded enough for the baby to develop normally. Verimetra, Inc., Carnegie Mellon University, and Children's Hospital of Pittsburgh have embedded a MEMS flow sensor and a series of micromachined bar codes on the tip of a balloon catheter. The bar codes enable the catheter to be visualized with ultrasound allowing surgeons to know its exact position. The flow sensor is a thermistor. As blood flow increases the temperature measured by the thermistor decreases and in this manner blood flow changes as the catheter progresses through the heart's vessels can be monitored. This allows for the measurement of blood flow at or near a constriction and then again after the procedure to open the constriction has been performed.

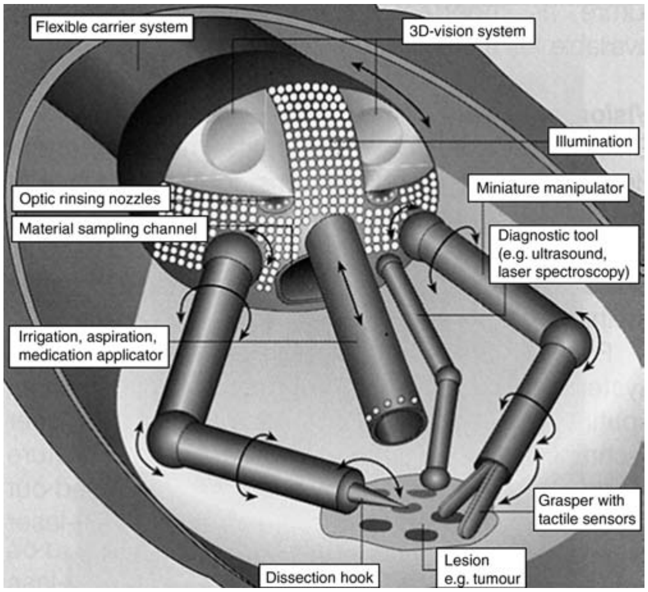


Figure 14. Highly integrated microsurgical probe. (Courtesy of Prof. Dr. M. Schurr, novineon Healthcare Technology Partners GmbH and Prof. G. Buess, Universitätsklinikums Tbingen.)

FUTURE SYSTEMS

In 1959, Richard Feynman gave his famous talk There's Plenty of Room at the Bottom (35). In it, he talked about being able to put a surgeon in a blood vessel which would be able to look around ones heart. Initially, new microsurgical tools will focus on measuring or detecting a specific parameter be it pressure, blood flow, velocity, temperature, and so on. The MEMS sensor systems will continue to be refined and improved leading to the integration of multiple sensor systems on surgical tool followed by tools with multiple functions. Multifunction surgical tools will reduce the number of tool insertions and removals reducing patient risks. Eventually, this will lead to highly integrated probes, which will do everything a surgeon needs, such as the concept shown in Fig. 14 (36). These tools will fit through a standard 5-mm port and have built in 3D cameras for visualization, biopsy samplers with microfluidic processing capability to do tissue analysis, ultrasound transducers, and tactile sensors for feedback to the surgeon.

BIBLIOGRAPHY

1. Diamiano R. Next up: Surgery by remote control. NY Times, Apr. 4, 2000; pp. D1.
2. Blazer E, Koh W, Yon E. A miniature digital pressure transducer. Proceedings of the 24th Annual Conference on Engineering Medicine and Biology. 1971. pp 211.
3. Peirs J, et al. A microturbine made by micro-electro-discharge machining. Proc 16th Eur Conf Solid-State Transducers 2002; 790–793.
4. Pique A, et al. Laser direct-write of embedded electronic components and circuits. Presented at Photon Processing

in Microelectronics and Photonics IV, Jan 24–27 2005, San Jose, (CA); 2005.

5. Götzen R. Growing up, additive processes in MEMS fabrication and packaging. Presented at Machines and Processes for Micro-scale and Meso-scale Fabrication, Metrology and Assembly, Gainesville, FL; 2003.
6. Rebello KJ, Lebouitz K, Migliuolo M. MEMS tactile sensors for surgical instruments. MRS Symp. Proc.: Biomicroelectromech. Systems (BioMEMS). 2003;773:55–60.
7. Menciassi A, et al. Force feedback-based microinstrument for measuring tissue properties and pulse in microsurgery. Presented at 2001 IEEE International Conference on Robotics and Automation (ICRA), May 21–26 2001, Seoul; 2001.
8. Rebello KJ, et al. MEMS based technology for endoscopic assessment of coronary arterial hardness. Presented at the 5th Annual NewEra Cardiac Care Conference, Dana Point (CA); 2002.
9. Bicchi A, et al. Sensorized minimally invasive surgery tool for detecting tissutal elastic properties. Presented at Proceedings of the 1996 13th IEEE International Conference on Robotics and Automation. Pt. 1 (of 4), Apr. 22–28 1996, Minneapolis, (MN); 1996.
10. Rosen J, Hannaford B, MacFarlane MP, Sinanan MN. Force controlled and teleoperated endoscopic grasper for minimally invasive surgery—experimental performance evaluation. IEEE Trans Biomed Eng 1999;46:1212–1221.
11. Scilingo EP, Bicchi A, De Rossi D, Iacconi P. Haptic display able to replicate the rheological behaviour of surgical tissues. Presented at Proceedings of the 1998 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Pt. 4 (of 6), Oct. 29-Nov. 1 1998, Hong Kong, China; 1998.
12. Fleischman A. et al. Miniature high frequency focused ultrasonic transducers for minimally invasive imaging procedures. Sensors Actuators, A: Phys 2003;103:76–82.
13. Zara JM, Bobbio SM, Goodwin-Johansson S, Smith SW. Intracardiac ultrasound scanner using a micromachine (MEMS) actuator. IEEE Trans Ultrasonics, Ferroelectrics, and Frequency Control 2000;47:984–993.
14. Daft C, et al. Microfabricated ultrasonic transducers monolithically integrated with high voltage electronics. Presented at 2004 IEEE Ultrasonics Symposium, Aug. 23–27 2004, Montreal, Queeue, Canada; 2004.
15. Chang JK, et al. Development of endovascular microtools. J Micromech Microeng 2002;12:824–831.
16. Degertekin FL, Guldiken RO, Karaman M. Micromachined capacitive transducer arrays for intravascular ultrasound. Presented at MOEMS Display and Imaging Systems III, Jan. 24–25 2005, San Jose, (CA); 2005.
17. Brock D. Smart medicine: The application of Auto-ID technology to healthcare. Auto-ID Center, MIT, Cambridge, (MA); 2002.
18. Tatar F, Mollinger JR, Bastemeijer J, Bossche A. Time of flight technique used for measuring position and orientation of laparoscopic surgery tools. Presented at Sensors, 2004. Proceedings of IEEE; 2004.
19. Fang C-M, Lee S-C. A research of robotic surgery technique by the use of MEMS accelerometer. Presented at Engineering in Medicine and Biology, 2002. 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society. EMBS/BMES Conference; 2002. Proceedings of the Second Joint, 2002.
20. Tanase D, et al. 3D position and orientation measurements with a magnetic sensor for use in vascular interventions. Presented at Biomedical Engineering, 2003. IEEE EMBS Asian-Pacific Conference on; 2003.
21. Totsu K, Haga Y, Esashi M. Three-axis magneto-impedance effect sensor system for detecting position and orientation of catheter tip. Sensors Actuators, A: Phys 2004;111:304–309.
22. Chen X, Lal A. Integrated pressure and flow sensor in silicon-based ultrasonic surgical actuator. Presented at Ultrasonics Symposium, 2001 IEEE; 2001.
23. Son I-S, Lal A. Feedback control of high-intensity silicon ultrasonic surgical actuator. Presented at Solid-State Sensors, Actuators and Microsystems, 2005. Digest of Technical Papers. TRANSDUCERS '05. The 13th International Conference on; 2005.
24. Polla DL, et al. Microdevices in medicine. Annu Rev Biomed Eng 2000;2:551–76.
25. Polla D, et al. Precision micromotor for surgery. Presented at Microtechnologies in Medicine and Biology, 1st Annual International, Conference On. 2000; 2000.
26. Kozlova TV, Shaposhnikova NF, Scobeleva VB, Sokolovskaya TV. Non-penetrating deep sclerectomy: Evolution of the method and prospects for development (review). Ophthalmosurgery 2000;3:39–54.
27. Angunawela R, Von Mohrenfels CW, Marshall J. A new age of cataract surgery. Cataract & Refractive Surgery Today I 2005; 36–38.
28. Kalvesten JE, Smith L, Tenerz L, Stemme G. First surface micromachined pressure sensor for cardiovascular pressure measurements. Presented at Proceedings of the 1998 IEEE 11th Annual International Workshop on Micro Electro Mechanical Systems, Jan. 25–29 1998, Heidelberg, Ger; 1998.
29. Tanase D, Goosen JFL, Trimp PJ, French PJ. Multi-parameter sensor system with intravascular navigation for catheter/guide wire application. Presented at Transducers'01 Eurosensors XV, Jun. 10–14 2001; Munich, 2002.
30. Haga Y, Esashi M. Biomedical microsystems for minimally invasive diagnosis and treatment. Proc IEEE Biomed App Mems Microfluidics 2004;92:98–114.
31. Kathuria YP. An overview on laser microfabrication of biocompatible metallic stent for medical therapy. Presented at Laser-Assisted Micro- and Nanotechnologies 2003, Jun. 29–Jul. 3 2003, St. Petersburg, Russian Federation; 2004.
32. Reed ML. Micromechanical systems for intravascular drug and gene delivery. Presented at BioMEMS 2002 Conference, Boston; 2002.
33. Goldschmidt-Clermont P, Kandzari D, Khouri S, Ferrari M. Nanotechnology needs for cardiovascular sciences. Biomed Microdevices 2001;3:83–88.
34. Ruzzu A, et al. A cutter with rotational-speed dependent diameter for interventional catheter systems. Presented at Micro Electro Mechanical Systems, 1998. MEMS 98. Proceedings., The Eleventh Annual International Workshop on, 1998.
35. Feynman RP. There's plenty of room at the bottom. J Microelectromech Systs 1992;1:60–66.
36. Schurr MO, Heyn S-P, Ment W, Buess G. Endosystems—Future perspectives for endoluminal therapy. Minimally Invasive Ther Allied Technol 1998;7:37–42.

Further Reading

Taylor RH, Lavallee S, Burdea GC, Mosges R. Computer Integrated Surgery: Technology and Clinical Application. Cambridge, (MA): MIT Press; 1996.

Zenati M. Robotic heart surgery. Cardiol Rev 2001;9(5):1–8.

Davies B. A review of robotics in surgery. Proc Inst Mech Eng 2000;214:129–140.

Madou M. Fundamentals of Microfabrication: The Science of Miniturization. 2nd ed. Boca Raton, (FL): CRC Press; 2002.

Kovacs GTA. Micromachined Transducers Sourcebook. Boston, MA: McGraw-Hill; 2002.

See also Endoscopes; fiber optics in medicine; intrauterine surgical techniques; nanoparticles; radiosurgery, stereotactic.

MINIMALLY INVASIVE SURGICAL TECHNOLOGY

Jay R. Goldberg
Marquette University
Milwaukee, Wisconsin

INTRODUCTION

Most surgical procedures involve the invasion and disruption of body tissues and structures by surgical instrumentation and/or implantable medical devices, resulting in trauma to the patient. Diagnostic imaging procedures, such as magnetic resonance imaging (MRI), computed tomography (CT), X ray, positron emission tomography (PET), and ultrasound do not require disruption of body tissues, and are thus considered to be noninvasive. Extra corporeal shockwave lithotripsy (ESWL) used to disintegrate kidney stones is an example of a noninvasive therapeutic procedure. As shown in Fig. 1, it focuses acoustic energy, generated outside of the patient's body, on kidney stones. No trauma to the patient occurs during this procedure.

Open heart coronary artery bypass and organ transplantation surgery are examples of highly invasive surgery. These procedures require a high level of invasion and disruption of body tissues and structures. Bones, muscle tissue, and blood vessels are cut, and tissue from other parts of the body may be grafted, resulting in a high level of surgical trauma to the patient.

Minimally invasive surgical procedures are less traumatic than corresponding conventional surgical procedures. The use of small instruments placed through intact natural channels, such as the esophagus, urethra, and rectum, is less invasive than conventional open surgical approaches requiring large incisions, significant loss of blood, and trauma to tissues. The use of small instruments, such as biopsy guns and angioplasty balloons placed through small incisions, results in minor trauma to the patient, but is much less invasive than open surgical procedures used to accomplish the same goals. Procedures using small instruments through intact natural channels or small incisions are classified as minimally invasive.

A minimally invasive surgical procedure can be defined as surgery that produces less patient trauma and disruption of body tissues than its conventional open surgical counterpart. For example, conventional appendectomies require a 4 cm long incision made through layers of skin, muscle, and other tissues to gain access to the appendix. Once the appendix is removed, the layers of the wound are sutured together to allow them to heal. This invasive procedure requires a significant level of invasion and disruption of tissues and other structures. The minimally invasive appendectomy is performed with small surgical instruments placed into small incisions in the patient's abdomen. Once the appendix is removed, the small incision is closed with sutures. This procedure results in much less trauma to the patient than the open surgical approach.

Minimally invasive surgery (MIS) is performed with small devices inserted through intact natural orifices or channels, or small incisions used to create an orifice. Some procedures are performed with devices located outside the body, and thus are noninvasive. The MIS procedures are an alternative to open surgery. The benefits of MIS include reduced patient trauma, postoperative recovery time, and healthcare costs.

INSTRUMENTATION FOR MINIMALLY INVASIVE SURGERY

Many MIS procedures involve flexible or rigid fiber optic endoscopes for imaging surgical sites and delivering instrumentation for diagnostic or therapeutic applications. The term "endoscope" is a generic term used to describe tubular fiber optic devices placed into the body to allow visualization of anatomical structures.

Endoscopes consist of a fiber optic light guide, high intensity light source, coherent fiber optic bundle, steering mechanism, and various working channels for insertion of endoscopic instrumentation and infusion of irrigation fluids or insufflating gases. Glass fibers comprise the fiber optic light guide and coherent bundle and are surrounded by a flexible polyurethane sheath or rigid stainless steel tube. Figure 2 shows the components of a typical endoscope. The light source provides light that is transmitted through the light guide and projected onto the anatomical site to create an image. The coherent fiber bundle transmits the image back to the focusing eyepiece and into the surgeon's eye. The surgeon can move the endoscope proximally (toward the patient) and distally (away from the patient) by pushing and pulling the endoscope into and out of the body, respectively. Steering is accomplished with a handle attached to a cable that pulls and bends the tip of the endoscope in the desired direction. The proximal (closest to the patient) end of the endoscope (shown in Fig. 3) contains lumens for the fiber optic light guide and coherent fiber bundle, and one or more working channels for passage of instruments and irrigation fluid into and out of the operative site. Some endoscopes contain video cameras that display endoscopic images on a video monitor in the operating room, as shown in Fig. 4.

Specialized endoscopes have different names depending on what part of the body they are used to image. Table 1 lists the names of various endoscopes, the procedures for which they are used, and the anatomical location where they are used.

Endoscopic Procedures and Instrumentation

To perform an endoscopic procedure, the surgeon inserts a sterilized endoscope into a natural channel or orifice, such as the urethra, rectum, esophagus, bronchial tube, or nose. If there is no natural channel to provide access to the operative site (as in abdominal surgery), then the surgeon will create one with a small incision, and may insert a hollow trocar into the incision, as shown in Fig. 5. The trocar is left in the wound to provide access to the abdominal cavity. The endoscope is inserted into the access site (natural channel, incision, or trocar) and as it is advanced toward the operative site the surgeon monitors the image produced by the endoscope, either through the objective eyepiece or video monitor (as shown in Fig. 4). Rigid endoscopes are typically used when access is obtained

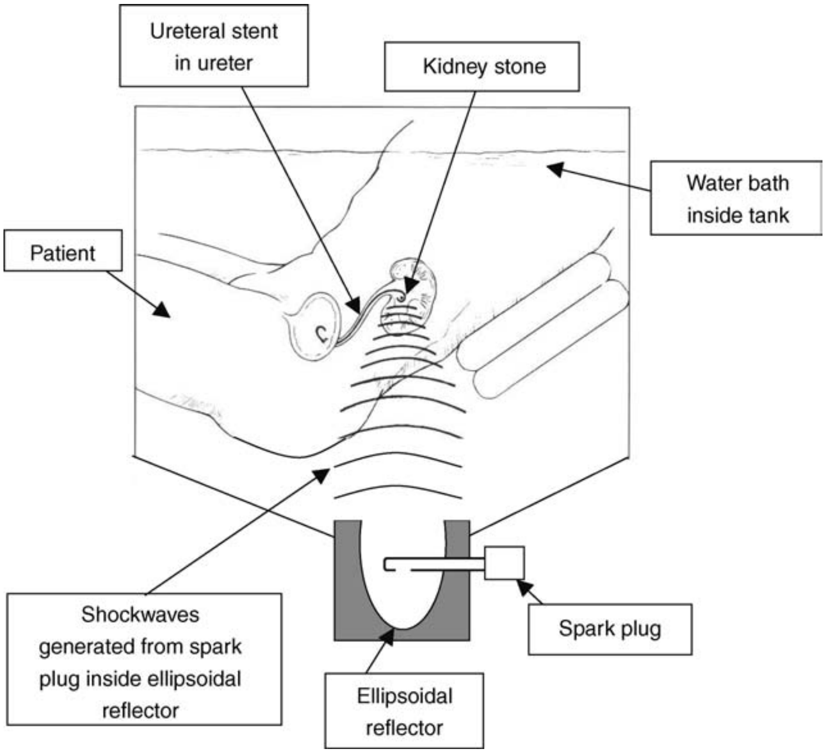


Figure 1. Schematic diagram of noninvasive extra corporeal shockwave lithotripsy (ESWL) procedure used to disintegrate kidney or biliary stones. Acoustic waves generated from spark plugs are focused with ellipsoidal reflectors on the kidney stone.

through an incision. Flexible endoscopes are used when access is obtained through natural channels (2). Once in position, the surgeon can then manipulate the endoscope to inspect tissues and anatomical structures (Table 1).

If a procedure other than visual inspection is required, the surgeon has a variety of instruments available to grasp, cut, remove, suture, and cauterize tissues, and remove debris through the endoscope. Forceps (Fig. 6) and graspers are used to grasp tissue and other objects such as kidney stones. Scalpels and scissors (Fig. 7) are used for cutting tissue. Suturing devices are used to repair internal wounds resulting from endoscopic procedures such as appendectomies, cholecystectomies (gallbladder removal), and arthroscopies. Morcellators are used to reduce the size and change the shape of a mass of tissue, such as a gallbladder, allowing it to be removed through a small incision. Electrohydraulic lithotriptor (EHL) and laser probes are used to disintegrate objects, such as kidney

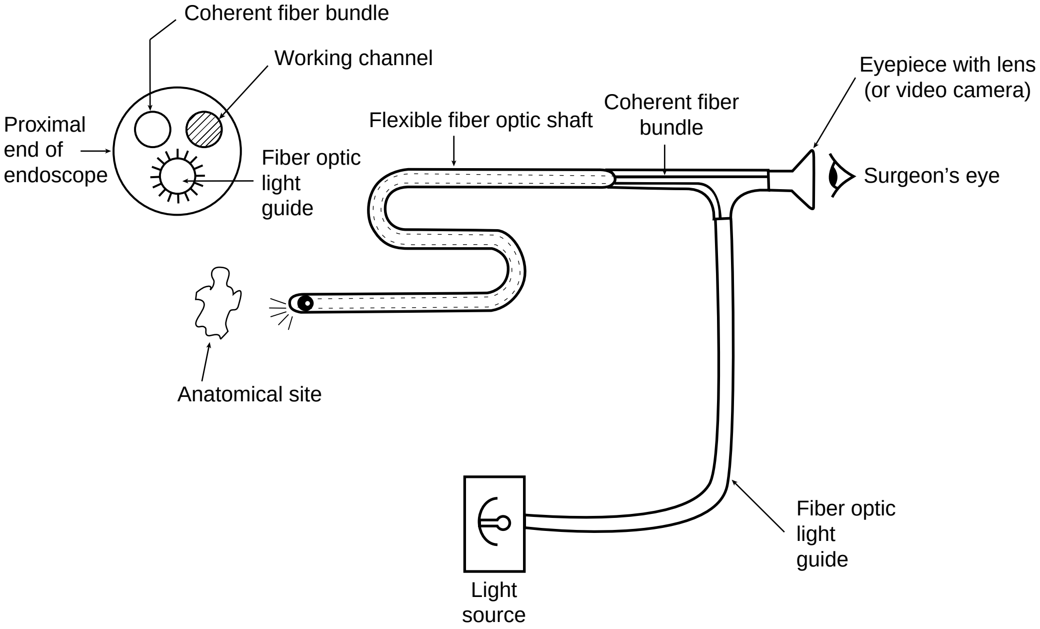


Figure 2. Components of a typical endoscope.

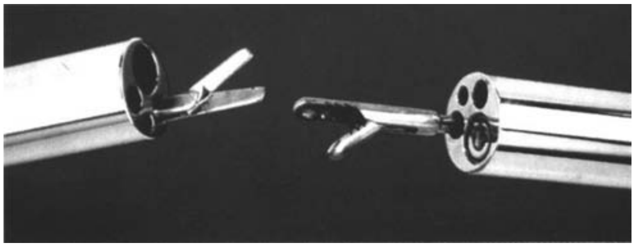


Figure 3. Proximal ends of typical endoscopes. Endoscopic instruments (left:scissors, right:forceps) are shown placed through working channels. Two other channels contain the fiber optic light guide and coherent fiber bundle, respectively. A fourth channel can be used for instrumentation or delivery of irrigation fluids to operative site (1). (Reprinted from Endoscopic Surgery, Ball, K., page 62, Copyright 1997, with permission from Elsevier.)

stones, with acoustic and laser energy, respectively, allowing the debris to be passed through the ureter. Stone baskets are used to trap kidney or ureteral stones for endoscopic removal, as shown in Fig. 8. These instruments are placed through the working channel of an endoscope or trocar and are operated by manipulating handles and controls located outside the patient.

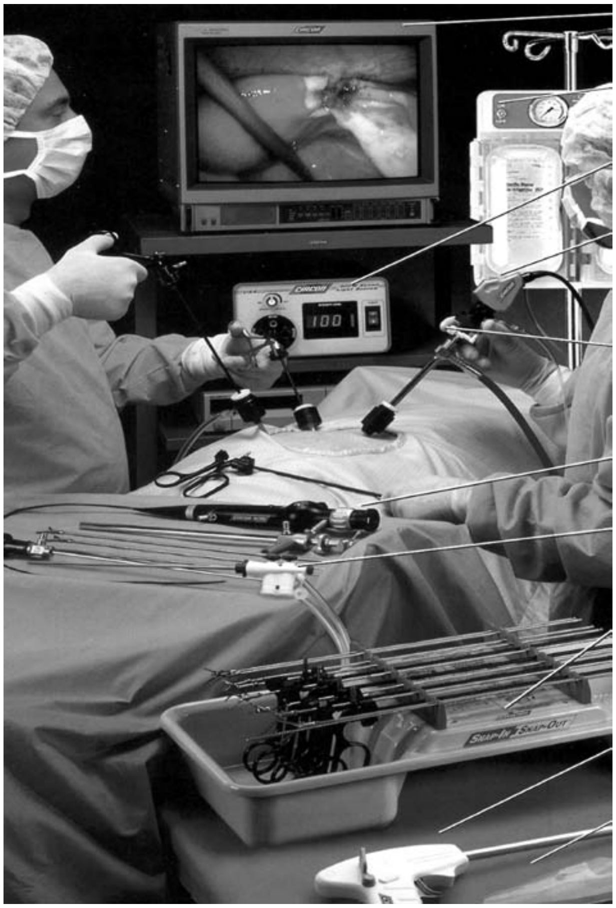


Figure 4. Surgeon viewing laparoscopic images of gallbladder on video monitor in operating room. Note use of three trocars in patient's abdomen; one each for video camera and two instruments. (Courtesy ACMI, Southborough, MA.)

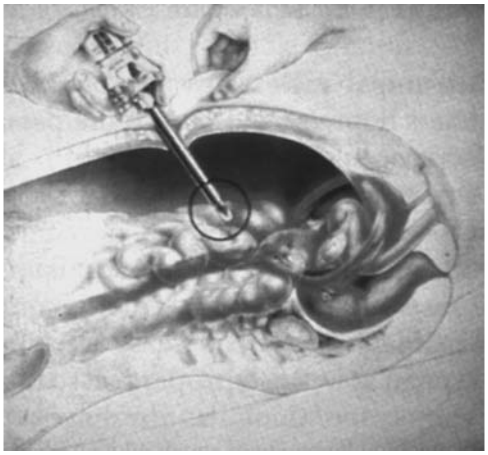


Figure 5. Trocar inserted into abdominal incision to provide access to abdominal cavity. Safety shield covers sharp end of trocar (circled) upon insertion to prevent damage to abdominal organs. (1) (Courtesy Ethicon Endo-Surgery Inc., Cincinnati, OH.)

Endoscopes may contain multiple working channels that allow for irrigation of fluids to irrigate and flush out clots and other surgical debris. The presence of the fluid improves visibility through the endoscope by keeping the visual field clear and the lens clean. Laparoscopic surgery requires insufflation of CO_2 or N_2O through the working channel or trocar and into the abdominal cavity to cause the cavity to expand, separating organs, and enlarging the operative site and visual field.

There are typically not enough working channels in a laparoscope to accommodate all of the instruments needed for a particular procedure. In these cases, multiple access sites are created with additional trocars. Typically, a camera is placed through one trocar and used to visualize the work being performed with instruments placed through other trocars.

When an endoscopic procedure is completed, the endoscope and instrumentation are removed from the patient via the natural channel or incision. When a laparoscopic procedure is completed, the laparoscope, camera, instruments, and trocars are removed from the patient. The wounds created by the trocars are sutured and the patient begins the recovery process. Although some of the CO_2 from insufflation may escape the abdominal cavity when all instrumentation is removed, some will be absorbed by body tissues and eliminated via respiration. Patients typically recover and are discharged from the hospital within 24–48 h.

Non-Endoscopic Procedures and Instrumentation

Not all minimally invasive procedures require endoscopic devices for imaging and placement of instruments. Some MIS procedures (listed in Table 2), such as stereotactic radiosurgery, use lasers or gamma radiation in place of scalpels. Others use small catheters to deliver medication, devices, or energy to specific anatomical locations. Balloons and stents, delivered via catheters, are used to access and dilate constricted vessels and maintain patency. Catheter mounted electrodes are used to deliver thermal, micro-

Table 1. Names of Endoscopes, MIS Procedures with Which they are Used, and Anatomical Location Where they are Used

Medical Specialty	Type of Endoscope Used	MIS Procedures	Anatomical Location
Urology	Cystoscope Ureteroscope Nephroscope	Cystoscopy, Transurethral resection of the prostate (TURP) Ureteroscopy, stone removal Nephroscopy, stone removal	Urethra, bladder ureter, kidney
Gastroenterology	Gastroscope Colonoscope Sigmoidoscope	Gastroscopy, gastric bypass Colonoscopy, Sigmoidoscopy	Stomach, colon Sigmoid colon
General surgery	Laparoscope	Laparoscopy, hernia repair, appendectomy, cholecystectomy (gallbladder removal)	Abdomen
Orthopedics	Arthroscope	Arthroscopy	Knee and other joints
Ob/Gyn	Hysterescope	Tubal ligation, hysterectomy	Female reproductive tract
Ear, nose, and throat	Laryngoscope, Bronchoscope Rhinoscope	Laryngoscopy, bronchoscopy Rhinoscopy, sinuscopy	Larynx, bronchus, nose, sinus cavities

wave, or radio frequency energy to selectively destroy tissue in ablation procedures used to treat cardiac arrhythmias (3), benign prostatic hypertrophy (BPH) (4), and other conditions. Many of these devices are guided through the body with the help of imaging or surgical navigation equipment.

Image Guided Surgery–Surgical Navigation

Image guided surgery (IGS) allows surgeons to perform minimally invasive procedures by guiding the advancement of instrumentation through the patient's body with increased accuracy and better clinical outcomes. Preoperatively, an IGS system is used to produce computerized anatomical maps of the surgical site from MRI or CT images of the patient. These maps are then used to plan the safest, least invasive path to the site. During an image guided procedure, the IGS system provides surgeons with a three dimensional image showing the location of instruments relative to the patient's anatomical structures. It tracks the movement of surgical instruments in the body, correlates these movements with the patient's preoperative images, and displays the location of the instruments on a monitor in the operating room. This feedback helps the surgeon safely and accurately guide instruments to the surgical site, reducing the risk of damaging healthy tissue (4,5).

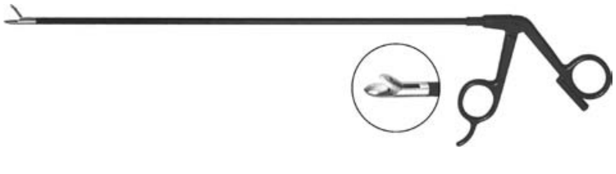


Figure 6. Endoscopic forceps. (Courtesy ACMI, Southborough, MA.)

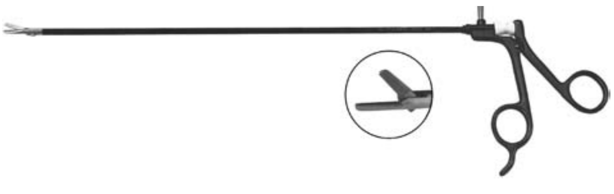


Figure 7. Endoscopic scissors. (Courtesy ACMI, Southborough, MA.)

An IGS system includes a computer workstation, localization system, display monitor, and specialized surgical instruments capable of being tracked by the system. Image processing and surgical planning software are also used. Tracking of instruments is accomplished through optical, electromagnetic, or mechanical means. Optical tracking systems use a camera mounted to view the surgical field and optical sensors attached to surgical instruments. These systems required line of sight between the camera and sensor to function properly. Electromagnetic tracking systems include a transmitter located close to the surgical site, and receivers attached to surgical instruments. Line of sight is not an issue with these systems, however, nearby metallic objects may produce interference to signals used to track instruments (4).

To ensure that the patient's anatomical features (and location of instruments) are accurately displayed by the IGS system, actual anatomical locations must be registered to the preoperative images. This can be accomplished by touching a probe to a marker on the patient's body and then assigning the location of this marker to its corresponding point in preoperative images (4).

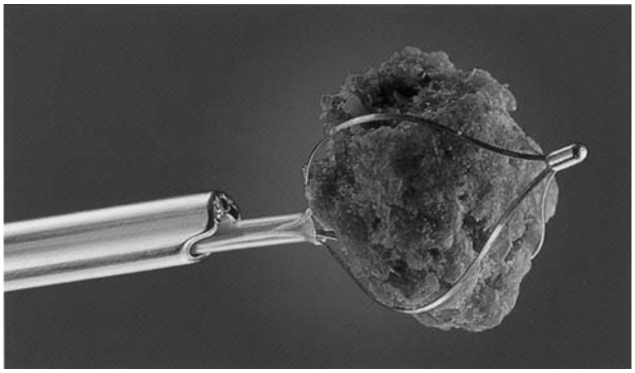


Figure 8. A stone basket used to trap and remove a large ureteral stone. (Courtesy ACMI, Southborough, MA.)

Table 2. Examples of Non-Endoscopic Minimally Invasive Surgical Procedures

Medical Specialty	Non-Endoscopic Minimally Invasive Surgical Procedure
Cardiovascular surgery	Minimally invasive direct coronary artery bypass (MIDCAB)
	Percutaneous transluminal coronary angioplasty (PTCA)
	Coronary stenting
	Radio frequency cardiac ablation
	Laser angioplasty
	Microwave catheter ablation for arrhythmias
	Chemical ablation for ventricular tachycardia
Ophthalmology	Laser photorefractive keratotomy (PRK)
	Laser ablation of cataracts
	Laser glaucoma surgery
Orthopedics	Total joint arthroplasty (hips, knees, and others)
Neurosurgery	Stereotactic radiosurgery
	Stereotactic radiotherapy
	Laser ablation of brain tissue, tumor tissue
Radiology	Clot removal
	Aneurism repair
	Cerebral arterial venous malformation repair
	Transjugular hepatic portal systemic shunt creation

Most surgical instruments must be adapted for use in image guided surgery by mounting sensors and other devices to allow detection of the instrument's position by the IGS system. Some medical device manufacturers are developing navigation-ready surgical instruments that contain small reflective spheres to act as reference arrays for cameras used in image guided surgery (5).

MINIMALLY INVASIVE SURGICAL PROCEDURES

This section contains a few examples of minimally invasive surgical procedures used in urology, orthopedics, neurosurgery, and general and cardiovascular surgery.

Ureteroscopy

Flexible ureteroscopy is used to perform a variety of diagnostic and therapeutic urological procedures. It involves entry into the body via intact natural channels (urethra and ureter) and does not require an incision. Local anesthesia and sedation of the patient are required.

Ureteroscopy is commonly used to remove ureteral or kidney stones. Initially, a cystoscope is inserted into the urethra. Next, a guidewire is placed through a cystoscope into the ureter, and advanced up into the kidney. While maintaining the position of the guidewire, the cystoscope is removed, and the distal end of the guidewire is placed into a working channel at the proximal end of the ureteroscope. The ureteroscope is then advanced along the guidewire into the ureter or kidney, as shown in Fig. 9. Active (controlled by the cable and handle) and passive deflection of the shaft, along with rotation of the flexible ureteroscope allows visual inspection of the renal calices as shown in Fig. 10. Ureteroscopic instruments are then placed through the working channel into the ureter or kidney. Figure 11 shows two devices used for ureteroscopic removal of ureteral stones. The stone grasper is used to physically grab the stone (Fig. 11). If the stone is small enough to fit into the working channel, it is pulled out of the patient through the ureteroscope. Large stones that will not fit into the working channel are pulled out with the ureteroscope. The laser lithotripter (Fig. 11) disintegrates the stone into small particles that can easily be passed naturally through the ureter, bladder, and urethra. Collapsed stone baskets can be placed alongside a stone and moved proximally and distally as they are expanded, until the stone is trapped in the basket (as shown in Fig. 8) and pulled out of the urethra.

Laparoscopy

Laparoscopy is commonly used for removal of the gallbladder and appendix, hernia repair, and other abdominal procedures. The basic steps involved in laparoscopy have been previously described. The lack of natural channels located in the abdomen requires the use of trocars to gain access to the operative site. Laparoscopic procedures require insufflation of gas to separate organs and expand the visual field. This is controlled by a separate insufflator that controls the pressure inside the abdomen produced by the flow of insufflating gases (1).

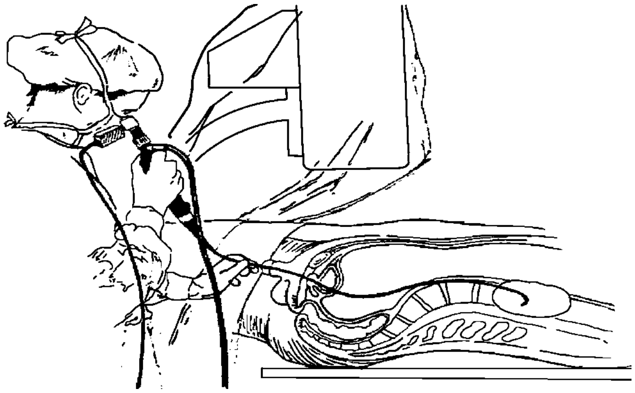


Figure 9. A surgeon views images through a ureteroscope placed through the urethra, bladder, ureter, and into the kidney. (Courtesy ACMI, Southborough, MA.)

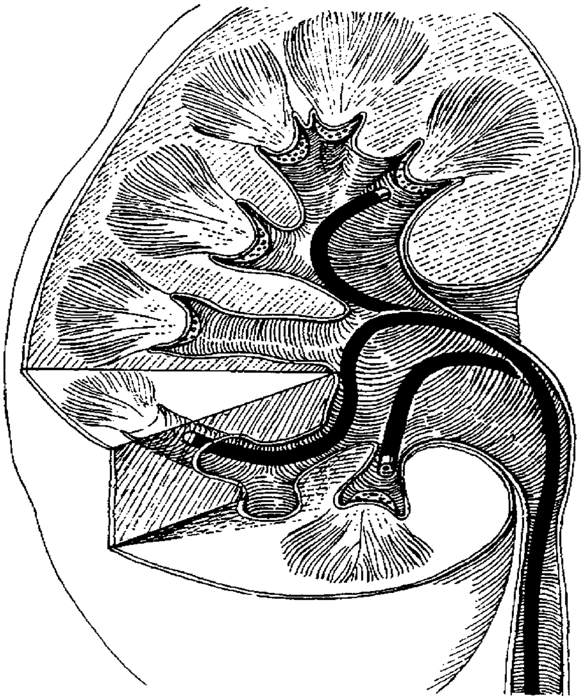


Figure 10. Flexibility and steerability of proximal end of ureteroscope allow inspection of upper, middle, and lower renal calices. (Courtesy ACMI, Southborough, MA.)

Total Joint Replacement

Total hip and knee replacements are typically performed with large incisions to expose the joint, allowing for complete visualization of and access to the joint and soft tissues. New surgical approaches using smaller incisions that result in less damage to muscles and other soft tissue limit the view of the joint. To compensate for the limited view, fluoroscopy and IGS systems are often used. Some existing surgical instruments have been modified to enable surgery through smaller incisions (6).

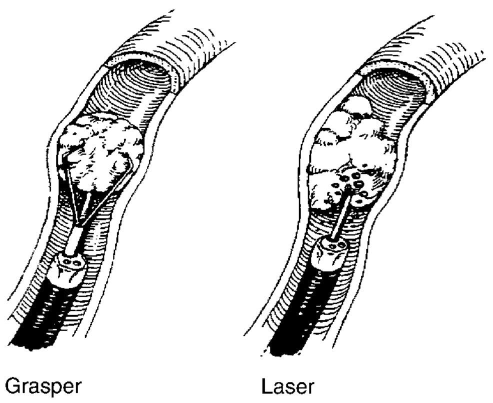


Figure 11. Devices used for endoscopic removal of ureteral stones. (Courtesy ACMI, Southborough, MA.)

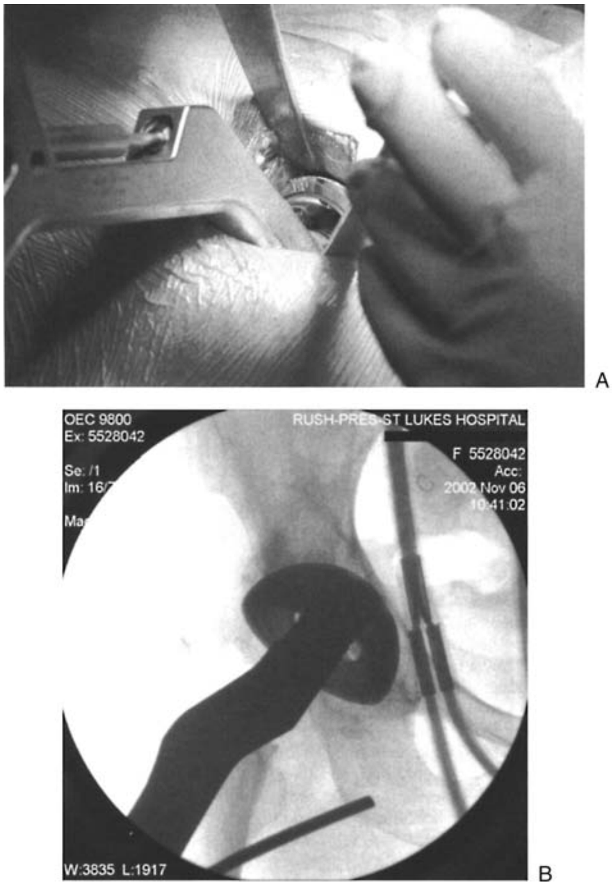


Figure 12. Acetabular component inserted with inserter through small incision (top image). Fluoroscopic image of inserter and acetabular component seated in acetabulum (bottom image). Images such as these allow surgeons to ensure proper placement of instruments and implantable components within hip joint when small incisions prevent viewing of entire device (6). (From MIS for the Hip and Knee: A Clinical Perspective, 2004, pg. 20, Minimally Invasive Total Hip Arthroplasty: The Two Incision Approach, Berger, R.A. and Hartzband, M.A., Fig. 2.11. Copyright 2004, with kind permission of Springer Science and Business Media.)

Traditional hip arthroplasty requires a 30–46 cm long incision, which is highly disruptive to muscles and surrounding tissues. Two different techniques can be used for minimally invasive total hip arthroplasty. One technique uses two 5 cm long incisions, one each for preparation and insertion of the acetabular and femoral components, respectively. The other technique involves one 8–10 cm long incision. Modified retractors and elevators are used to gain access and expose the joint. Fluoroscopy and IGS systems are used to provide the surgeon with a view of instruments and components (as shown in Fig. 12) and assist in positioning of instruments designed to enable accurate component alignment and placement. Minimally invasive hip arthroplasty results in less disruption of muscles and tissues, smaller and less noticeable scars, less blood loss and pain, and fewer blood clots and dislocations (6).

Most minimally invasive knee arthroplasties performed through smaller incisions involve a single compartment of the knee. These procedures typically use existing unicompartmental knee implants for resurfacing the medial or lateral femoral condyle and corresponding tibial plateau. Existing instrumentation has been modified to obtain access to the joint and enable accurate placement and alignment of the unicompartmental knee components through smaller incisions.

Minimally Invasive Direct Coronary Artery Bypass

Coronary arteries may become blocked due to fatty deposits (plaque), blood clots, or other causes. This reduces blood flow to cardiac muscle, depriving it of needed oxygen and other nutrients. This condition can result in a myocardial infarction (heart attack) that can damage cardiac muscle.

Traditional open heart coronary artery bypass graft surgery has been used to restore normal blood flow to cardiac muscle. This procedure involves grafting a new vessel to points on both sides of the blocked coronary artery, thereby bypassing the blockage and restoring normal blood flow. It requires open access to a still heart to allow suturing of the graft to the blocked coronary artery. A sternotomy and separation of the sternum is required to provide access to the heart. The heart is stopped and the patient is attached to a heart–lung machine to maintain circulation of oxygenated blood through the body during the surgical procedure. This procedure is highly invasive and can result in a variety of complications, many of which are associated with use of a heart–lung machine. Inflammatory responses negatively affecting multiple organ systems have been observed in patients who were perfused with a heart–lung machine during traditional open heart coronary bypass surgery (7). These responses are due to reactions between circulating blood and material surfaces present in the heart–lung machine.

Minimally invasive direct coronary artery bypass is an alternative to open heart surgery. A small 5–10 cm incision made between the ribs replaces the sternotomy to gain access to the heart. A retractor is used to separate the ribs above the heart to maximize access to the operative site (8). Heart positioners and stabilizers (Fig. 13) are used to minimize the motion of the heart, allowing surgeons to perform the procedure on a beating heart, eliminating the need for the heart–lung machine. Some stabilizers grasp the heart with suction cups. Others use a fork like device to apply pressure to the beating heart to keep it steady for anastomosis of the graft (8). A thorascope and small endoscopic instruments are used to visualize the surgical site and perform the surgical procedure. The left internal mammary artery is commonly used as the grafted

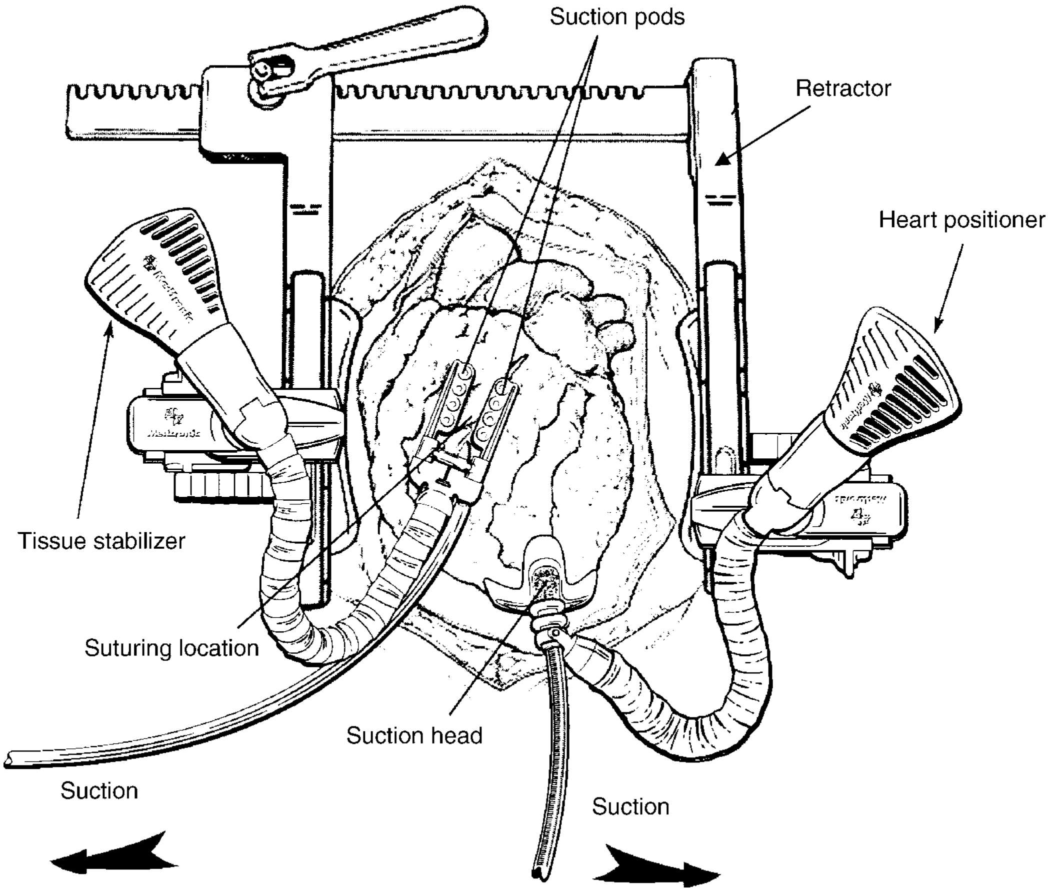


Figure 13. Heart positioner and stabilizer used in MIDCAB procedures. The positioner attaches to the sternal retractor and holds the heart in position using suction. It provides greater access to the blocked coronary artery. The tissue stabilizer, attached to the sternal retractor, straddles the blocked artery and holds the suturing site steady. This allows surgery on a beating heart, eliminating the need for a heart–lung machine along with its associated complications. (Courtesy of Medtronic, Inc., Minneapolis, MN.)

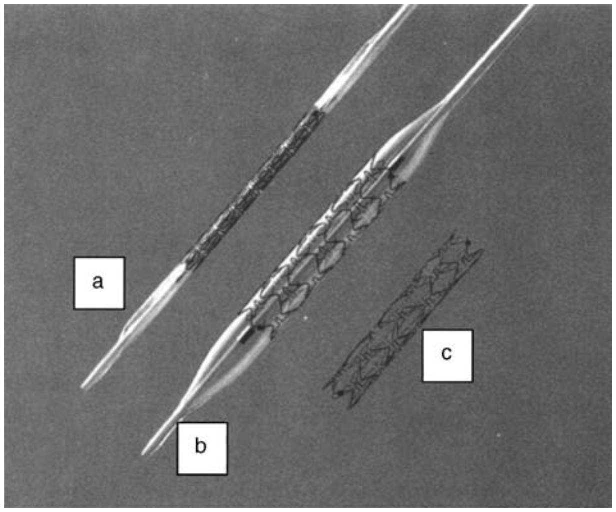


Figure 14. A PTCA catheter and stent. Stent and balloon in collapsed configuration for insertion and placement into coronary artery (a). Inflated balloon causing stent to expand (b). Expanded stent after collapse of balloon and removal of catheter (c). (Courtesy of Sorin Biomedica Cardio SpA, Italy.)

vessel to bypass the blockage of the coronary artery. The MIDCAB results in fewer complications, less blood loss, and shorter hospital stays.

Percutaneous Transluminal Coronary Angioplasty with Stent Placement

The PTCA method is used to open a blocked, constricted coronary artery instead of bypassing the blockage with a graft. Under local anesthesia, a steerable balloon catheter containing a stent mounted to the balloon (Fig. 14a) is inserted into the patient's femoral artery and guided to the constricted coronary artery under fluoroscopy. Once in position, the balloon is inflated to compress and flatten the plaque against the arterial wall, creating a larger opening for blood to flow through the coronary artery. The balloon is constructed with materials of different stiffness such that pressure from the inflating balloon is radially applied to the constricted area, and not to tissue outside of the constricted region (3). During balloon inflation, the stent is expanded to a larger diameter Fig. 14b, which is maintained after deflation of the balloon. The catheter is removed, and the expanded stent (Fig. 14c) is left in place to maintain a larger opening and prevent restenosis of the coronary artery.

Most coronary stents are made of stainless steel, nitinol, cobalt chrome molybdenum alloy, or gold (3). Some stents contain coatings to improve biological performance (biocompatibility and resistance to clot formation) and/or elute drugs into surrounding tissues to prevent restenosis of the coronary artery.

The PTCA method does not remove plaque from the coronary artery; it flattens it so it no longer restricts blood flow. Plaque removal methods involve lasers and rotational cutting devices (3).

Stereotactic Radiosurgery

In this noninvasive procedure, focused external beams of radiation are delivered to specific locations in the brain to treat tumors (9). The accuracy of the delivery prevents damage to healthy brain tissue. The patient's head is constrained in a mask or frame during the 30–45 min procedure.

Newer radiotherapy systems include robotic linear accelerators for delivery of radiation at any angle to the patient's head, a beam shaper to match the shape of the beam to the three-dimensional (3D) shape of the tumor, and imaging equipment to provide real-time tracking of tumor location and patient positioning (9).

Treatment of Benign Prostatic Hypertrophy

Benign prostatic hypertrophy causes the prostate to enlarge. An enlarged prostate applies pressure to the prostatic urethra that can occlude the urethra, reduce urinary flow rate, and make urination difficult. The enlarged prostate can be removed with an open surgical approach or less invasive transurethral resection of the prostate (TURP). The TURP procedure involves cutting and removing segments of the prostate through a cystoscope or with a resectoscope inserted into the urethra, and can result in complications, such as incontinence and impotence. Prostatic balloon dilators and prostatic stents inserted into the urethra have been used to expand the narrowed urethra. Transurethral laser incision of the prostate (TULIP) has been used to cut and remove prostate tissue. In this procedure, a catheter mounted laser facing radially outward toward the prostate delivers laser energy that cuts through the enlarged prostatic tissue, relieving pressure on the urethra.

Newer approaches to treating BPH include transurethral microwave therapy (TUMT) and transurethral needle ablation (TUNA) (4). These procedures use microwave and radio frequency energy, respectively, to heat and destroy unwanted tissue without actually removing the tissue. The TUMT method involves insertion of a urethral catheter containing a microwave antenna into the bladder neck. Channels contained in the catheter carry cooling water to prevent thermal damage to the urethra while microwave energy is used to heat the prostate for ~ 1 h. The TUNA method uses a cystoscope to insert two shielded electrode needles through the urethra and into the prostate. Radio frequency energy is delivered to the prostate, heating the tissue and destroying it. Thermal energy causes the prostate tissue to stiffen and shrink, relieving the pressure on the urethra caused by the enlarged prostate. Both TUMT and TUNA deliver controlled thermal energy to a targeted area to selectively destroy prostate tissue (4).

NEW DEVELOPMENTS IN TECHNOLOGY FOR MINIMALLY INVASIVE SURGERY

New devices and technologies are being developed and clinically evaluated for use in MIS. Robots can be used to enable surgeons to perform more intricate surgical procedures than can be done with endoscopic devices (10). In robot-assisted surgery, the surgeon operates a console to control mechanical arms positioned over the patient. The arms become an extension of the surgeon's hands. While observing the procedure on viewing screens (one for each eye) that produce a 3D image, the surgeon uses hand, finger, and foot controls to move the mechanical arms containing interchangeable surgical instruments through a small opening in the patient's body (10).

Other new technologies for MIS include 3D video imaging, integrated robotic and surgical navigation systems, devices for mechanical retraction of the abdominal wall (eliminating the need for insufflation), and telerobotics (8).

OUTCOMES OF MINIMALLY INVASIVE SURGERY

Comparison of MIS to Conventional Approaches

Minimally invasive surgical procedures result in reduced patient trauma, less postoperative pain and discomfort, and decreased complication rates. Hospital stays and postoperative recovery times are reduced, resulting in lower hospital costs and allowing patients to return to work earlier.

Studies comparing various MIS procedures to their traditional open surgical counterparts have been conducted. One such retrospective study compared the clinical and economic aspects of laparoscopic and conventional cholecystectomies (11). Results of 153 consecutive traditional procedures and 222 consecutive laparoscopic procedures performed in a German teaching hospital were evaluated. Researchers found that although laparoscopic cholecystectomies required longer operative times (92 vs. 62 min), they resulted in fewer complications (6 vs. 9), shorter hospital stays (3 vs. 8 days), and an 18% decrease in hospital costs, when compared to traditional procedures.

In a Canadian study, the direct costs of conventional cholecystectomy, laparoscopic cholecystectomy, and biliary lithotripsy were compared (12). Researchers concluded that laparoscopic cholecystectomy provided a small economic advantage over the other two procedures and attributed this to a shorter hospital stay.

In the United States, hospital stays following laparoscopic cholecystectomies typically ranged from 3 to 5 days. Now, many of these procedures are performed on an outpatient basis. This additional reduction in hospital stay further reduces hospital costs, resulting in a greater economic advantage over the conventional procedure.

A study comparing results of 125 consecutive off-pump coronary bypass (OPCAB) procedures to a matched, contemporaneous control group of 625 traditional coronary artery bypass graft (CABG) procedures was conducted (7). The OPCAB procedure is a beating heart procedure that does not involve a heart–lung machine. Partial sternotomies were used with some patients in the OPCAB group. Researchers found that the OPCAB procedure resulted in a lower mortality rate (0 vs. 1.4%), reduced hospital stays (3.3 vs. 5.5 days), 24% lower hospital costs, and a reduced transfusion rate (29.6 vs. 56.5%), when compared to the traditional CABG procedure. Excellent graft patency rates and clinical outcomes were also reported with the OPCAB procedure.

In another study of 67 MIDCAB patients, it was found that average hospital charges for MIDCAB were $14,676 compared to $22,817 for CABG and $15,000 for coronary stenting (13). The significantly lower charges for MIDCAB were attributed to shorter hospital stays, elimination of perfusion expenses, and reduction in ICU, ventilation, and rehabilitation times.

Limitations of Minimally Invasive Surgery

There are several problems and limitations associated with MIS procedures. First, some minimally invasive surgical procedures can take longer to perform than their more invasive counterparts (10). Second, open surgical procedures allow the surgeon to palpate tissue and digitally inspect for tumors, cysts, and other abnormalities. Tactile feedback also assists in locating anatomical structures. The inability of the surgeon to actually touch and feel tissue and structures at the operative site limits the diagnostic ability of some MIS procedures (10). Third, a two-dimensional (2D) image is produced by the endoscope and video monitor. The resulting loss of depth perception combined with restricted mobility of instruments through a small incision make manipulation of endoscopic instruments challenging. Fine hand movements are difficult due to the long distances between the surgeon's hand and working ends of these instruments. Fourth, there is a limit to the number of instruments that can be used at one time through trocars and working channels of a laparoscope. Finally, instrumentation for MIS procedures is more expensive.

Insufflation presents a small risk not associated with open surgical procedures (1,10). If the gas pressure inside the abdominal cavity becomes too high or there is a small defect in a blood vessel, then a gas bubble can enter the bloodstream and travel to the heart or brain, causing unconsciousness or death.

Laparoscopic removal of cancer tissue carries a very small risk of transporting cancer cells to other parts of the body. As tumor tissue is removed through the laparoscope, some cancer cells may remain in the body. They may be displaced from their original location resulting in the spread of cancer cells to other parts of the body (10).

SUMMARY

Minimally invasive surgery is made possible through the use of specialized devices and technologies. These include endoscopes, surgical instruments, imaging and navigation systems, catheters, energy delivery systems, and other devices. The MIS procedures tend to require more time, are more difficult to perform than conventional procedures, and present some unique risks not associated with conventional procedures. Compared to conventional open surgical procedures, MIS procedures demonstrate similar clinical efficacy and reduce trauma and disruption of body tissues and structures. Patients experience less pain and discomfort, fewer complications, and shorter recovery periods. Hospital stays are reduced, resulting in lower hospital costs.

BIBLIOGRAPHY

1. Ball K, Endoscopic Surgery. St. Louis: Mosby Year Book; 1997.
2. Holland P, The Fundamentals of Flexible Endoscopes, Biomedical Instrumentation and Technology. Association for the Advancement of Medical Instrumentation, p 343–348, Sep./Oct. 2001.
3. Webster J G, ed. Minimally Invasive Medical Technology. Institute of Phys Publishing Ltd., 2001.
4. Spera G, The Next Wave in Minimally Invasive Surgery, Medical Device and Diagnostic Industry, Canon Communications, August 1998.
5. Gill B, Navigation Surgery Changing Medical Device Development, Medical Device and Diagnostic Industry, Canon Communications, December 2004.
6. Scuderi R G, Tria A J, eds., MIS of the Hip and the Knee: A Clinical Perspective; New York: Springer-Verlag, 2004.
7. Puskas J, et al. Clinical outcomes and angiographic patency in 125 consecutive off-pump coronary bypass patients. Heart Surg Forum May 1999;2(3):216–221.
8. Spera G, The kindest cut of all, Medical Device and Diagnostic Industry, Canon Communications, July 1998.
9. Technology Trends: Radiation System Could Be an Alternative to Surgery, Biomedical Instrumentation and Technology, Association for the Advancement of Medical Instrumentation, p. 18, January/February 2005.
10. Comarow A, Tiny holes, big surgery. U.S. News & World Report, July 22, 2002.
11. Bosch F, Wehrman U, Saeger H, Kirch W. Laparoscopic or open cholecystectomy: Clinical and economic considerations. Eur J Surg 2002;168(5):270–277.
12. Conseil d'evaluation des technologies de la sante du Quebec (CETS). The Costs of Conventional Cholecystectomy, Laparoscopic Cholecystectomy, and Biliary Lithotripsy. Montreal: CETS, 1993.
13. Oz M, Goldstein D, Minimally Invasive Cardiac Surgery. Humana Press; 1999.

See also Endoscopes; gamma knife; microsurgery; stereotactic surgery; tissue ablation.

MOBILITY AIDS

Rory Cooper
Erik Wolf
Diane Collins
Eliana Chavez
Jon Pearlman
Amol Karmarkar
Rosemarie Cooper
University of Pittsburgh
Pittsburgh, Pennsylvania

INTRODUCTION

The National Center for Health Statistics estimates that 12.4 million adults are unable to walk a quarter of a mile in the United States, and that 28.3 million adults have moderate mobility difficulty (1). Approximately 4 million Americans use wheelchairs, and about one-half of them use their wheelchairs as their primary means of mobility. About 1.25 million people wear a prosthetic limb due to injuries, birth anomalies, and disease. Once fatal injuries are now survivable due to advancing medical achievements, prolonging life spans, and the staggering growth in the aging population. Because of this growth, the population of individuals who use mobility aids is sure to grow in the coming decades.

The goal of issuing wheelchairs, prosthetics, walkers or rollators to individuals with mobility impairments independence remains the top priority when prescribing one of these devices, other main concerns include safety, not causing secondary injury (i.e., pressure sores, carpal tunnel syndrome, rotator cuff tear), and the physical design of the device (e.g., weight, size, ease of use). Research has shown that manual wheelchair users commonly report shoulder, elbow, wrist, and hand pain (2). Low back pain and fatigue are also common secondary ailments experience due to exposure of whole-body vibrations (3). Safety research shows that proper fitting and wheelchair skill can reduce injurious tips and fall in wheelchairs (4).

In order to fully maintain an active lifestyle, including recreational activities, participating in the community, and going to work, transportation for people with mobility impairments is essential. With the added technology that is necessary to allow people to use public or private transportation, added safety features must also be included to maintain the security of both the drivers and the passengers.

Besides performing normal activities of daily living, sports, and recreational activity are an important physical and psychosocial aspect of any human being. The case is no different with people who use assitive technology. With dramatic advances in technology, people with mobility impairments are able to go places and participate in activities that were once nearly impossible.

In recent years, wheelchairs, prosthetics, walkers, and rollators have been designed to be stronger, safer, lighter, more adjustable, and smarter than in the past, through the use of materials like titanium and carbon fiber, using advanced electronics, and so on. The technology of wheelchairs, prosthetics, walkers, and rollators has improved dramatically in the past 25 years due largely in part to the increased demand of consumers, their loved ones and others who assist consumers, and the people who recommend and issue these technology devices. The term"recommend" is used because it is crucial that these devices, especially wheelchairs and prosthetics, be issued by a team of professionals to provide the highest levels of independence and safety, and that this team is centered around the client. All of these components are important to the further development of the technology and, in turn, they may result in the increased independence of people with mobility impairments.

CLINICAL ASSESSMENT OF MOBILITY

The ultimate goal and outcome for a clinical assessment of mobility should drive toward a successful wheelchair, prosthetics, walker, or rollator recommendation that enhances the quality of life expectations and their effectiveness as reported by the consumer. Quality of life is specific to and defined by each person and/or family receiving clinical services. The consumer, their family, and care givers, must be actively included in this process, as they will be most affected by the choice of the mobility aid. Also, people chose their mobility devices based on the features available that facilitate activities or address needs (5), and the clinician should be aware of the consumer preferences and the features of various devices.

The complexity of some of the mobility device components combined at times with involved disease processes can make it virtually impossible for a single clinician to act independently when recommending assistive technology. Therefore, involving *an interdisciplinary team* is recommended in the decision making process (6,7). This team, with the consumer as an involved member and a variety of rehabilitation professionals, includes a physiatrist or similarly trained physician who understands the importance of Assistive Technology, addresses medical issues and assists with mobility decisions; the Occupational or Physical Therapist with RESNA (www.resna.org) certified Assistive Technology Practitioner (ATP) credentials, who is the point person for evaluation and prescription; and the Rehabilitation Engineering (with RE Training and RET Credential) who is a technology expert with the ability to design–modify–tune devices, and who also understands the capabilities and applications of various technologies. Another important team partner is the Assistive Technology Supplier (ATS) or Certified Rehabilitation Supplier (CRTS, National Association of Rehabilitation Technology Suppliers, www.narts.org), who provides or manages demonstration equipment, does routine home and workplace assessments, and orders, assembles, and delivers the equipment. All team members involved in the mobility aid selection process should have knowledge about the technology available on the market. Peer reviewed journal articles (Assistive Technology, Archives of Physical Medicine and Rehabilitation, Journal of Rehabilitation Research and Development etc.), magazine articles and commercial database sources such as ABLEDATA (http://www.abledata.com/), or WheelchairNet (www.wheelchairnet.org) are good places to research devices or to direct consumers who want to inform and educate themselves.

Resources available to the team and its members includes a defined and dedicated space for demonstration equipment, assessments, and evaluations, access to common activities and tasks (e.g., ramps, curb cuts, bathroom, countertop), an electronic tracking system to follow clients and their technology, assessment resources (e.g., pressure mapping, SMART$^{\text{Wheel}}$, gait force plate, actigraph), IT Resources (e.g., email, web, databases, medline, paging, cell, wireless), and the facilities–hospital commitment to continuing education.

A mechanism for Quality Measures for AT Clinics will provide valuable feedback on performance quality and areas in need of improvement. An important tool to measure patient satisfaction is the information gained through a satisfaction survey provided to every patient, in order to find out whether the goals and desired outcomes have been met. Feedback on performance quality is provided through tracking mechanism of primary clinician credentials (ATS, ATP, RET), dedicated staffing levels, continuing education (CEUs and CMEs), compliance with the commission on Accreditation of Rehabilitation Facilities (CARF) AT Clinic Accreditation, as well as tracking of continuous quality improvement.

Assessment

The occupational or physical therapist conducts the initial assessment process and obtains critical information about the consumer and their environment. This part usually involves a structured interview with the consumer and then a physical motor assessment. Establishing a medical diagnosis that requires the mobility aid is vital to assure no ongoing medical problems exist that are not being adequately addressed. To properly specify a mobility device, the intentions and abilities of the consumer must be ascertained (8,9). The intentions and abilities may include how people perform tasks, where the deficits are, and how mobility systems can compensate for the deficits to augment task performance. Some outcome tools that are clinically used to measure the functional impact of mobility aids are the QUEST, the FEW and the Wheelchair Skills Test (WST).

Additional necessary information includes type of insurance, method of transportation, and physical capabilities. Also, if the consumer has been using a chair, historical information about their current chair should be addressed, including problems they are having. The mobility device chosen should also be compatible with the public and/or private transportation options available to the consumer, such as a bus, car, or van. The regularity of the surface, its firmness and stability are important, when, for example, recommending a wheelchair in determining the tire size, drive wheel location, and wheel diameter. The performance of a wheelchair is often dictated by the need to negotiate grades, as well as height transitions, such as thresholds and curbs. The clearance widths in the environment will determine the overall dimensions of the wheelchair. The climates the chair will be operated in, and the need to be able to operate in snow, rain, changing humidity and temperature levels, and other weather conditions, are important considerations as well.

A physical–motor assessment of strength, range of motion, coordination, balance, posture, tone, contractures, endurance, sitting posture, cognition, perception, and external orthoses is an important first step to obtain a basic understanding of an individual's capacity to perform different activities. The history likely provided significant insight related to their physical abilities. To verify this, a physical examination should focus on aspects of the consumer that (1) help justify the mobility aid, (2) help determine the most appropriate mobility aid, and (3) assure that medical issues are appropriately addressed.

Once the examination documents the need, or potential lack of need for the mobility device the remainder of the examination can focus on the appropriate technology. This is best assessed by giving the consumer the opportunity to try equipment to determine how functional and safe they maneuver/operate the device within the clinical space. During this part of the assessment, the consumer and family must be informed of the positive and negative characteristics of devices and services. The team needs to educate the consumer or family to participate in choosing the device that will meet their needs (Locus of Control) and assure the provision of follow-up services.

The in-home evaluation conducted by the ATS verifies that the device is compatible and will fit within the home environment of the consumer; that may also included recreational and work environment. Once the appropriateness of a device is established, final measurements are taken. For many people, a few simple measurements can be used to determine the proper dimensions for a wheelchair (10). Body measurements are typically made with the consumer in the seated position. A Rehabilitation Technology Supplier, therapist, or other member of the rehabilitation team often completes this.

MANUAL WHEELCHAIRS

When most individuals think of manual wheelchairs they envision the boxy, steel framed, standard wheelchairs commonly seen at airports and shopping malls. These wheelchairs may be acceptable for transport of short distances, but are good for little else. Their heavy weight and complete lack of adjustability makes them poor choices for anyone using a manual wheelchair for an extended period of time.

The development of the lightweight and ultralight wheelchairs evolved in the late 1970s with a select few, extremely motivated, manual wheelchair users choosing to perform modifications on their own wheelchairs to make them faster and easier to propel (11). After these initial steps the demand became far greater for lightweight, adjustable wheelchairs and several companies were created to meet that need.

Research conducted on the new lightweight and then ultralight manual wheelchairs have quantitatively shown the benefits over the heavier, nonadjustable standard style wheelchairs. Cooper et al. (12) reported that when subjected to the fatigue tests described in the ANSI/RESNA standards, ultralight manual wheelchairs were 2.3 times more cost effective than lightweight wheelchairs and 3.4 times more cost effective than depot style wheelchairs.

The benefits of lightweight and ultralight manual wheelchairs do not end with higher cost efficiency and longevity. They are also crucial in preserving the upper extremities of their users. Because of the constant use of the upper extremities by manual wheelchair users, they tend to experience secondary injuries such as joint pain, repetitive strain injury, and nerve damage (2). Compared to standard and lightweight wheelchairs, the ultralight wheelchairs have two very distinct advantages when attempting to prevent secondary injury in manual wheelchair users: the primary advantage is the lower weight (Fig. 1).

Standard style wheelchairs tend to be greater than 36 lb(16 kg), while lightweight wheelchairs tend to be $\sim$30–34 lb(14 + 18 kg) and ultralight wheelchairs $\sim$20–30 lb(9–14 kg). The second advantage presented by the ultralight wheelchairs is adjustability. Lightweight wheelchairs do have some adjustability, but not to the extent of an ultralight wheelchair. Masse et al. (13) showed that moving the horizontal axle position toward the front of the wheelchair and moving the seat downward created a more efficient position for wheelchair propulsion and resulted in less exertion without loss of speed. Although some studies have been conducted to asses the importance of wheelchair setup in reducing upper extremity injury in manual wheelchair users, the solution has not been clearly defined. However, what is certain is that the adjustability and weight of manual wheelchairs are crucial parameters when selecting a wheelchair for a user that will be propelling for extended periods of time.

Another recent addition to manual wheelchairs has been suspension elements, such as springs or dampeners to reduce the amounts of vibration transmitted to manual wheelchair users. During a normal day of activity for a wheelchair user, they encounter bumps, oscillations, and other obstacles that may subject them to whole-body vibration levels that are considered harmful. VanSickle et al. (3) demonstrated that when propelling over certain obstacles, vibrations experienced at the seat of the wheelchair and the head of the user exceed the safety levels prescribed by the ISO 2631-1 Standard for evaluation of human exposure to whole-body vibration (14). Wheelchair companies have attempted to reduce the amounts of transmitted vibration by adding suspension to manual wheelchairs. Research has been done to evaluate effectiveness of suspension at reducing vibration. Results show that on average, suspension does reduce the vibration levels, however, the designs are not yet optimally effective and may not be as effective based on the orientation of the suspension elements (15,16).

POWERED ASSIST WHEELCHAIRS

Often, people using manual wheelchairs are required to transition to using powered wheelchairs or are at a level of capacity where they must choose between the two. This may be because of increased amounts of upper extremity

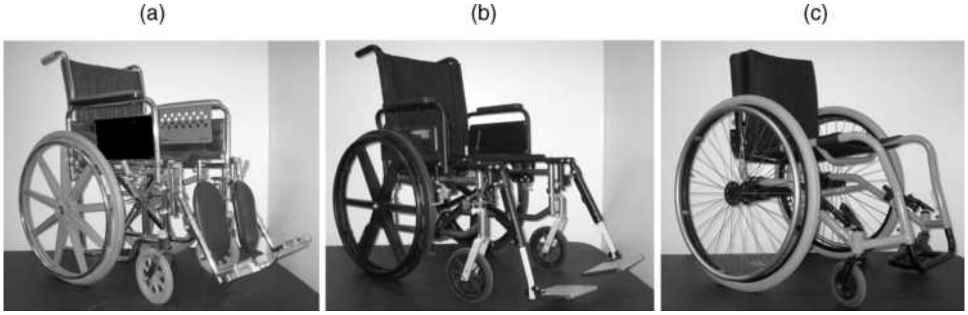


Figure 1. (a) Standard, (b) lightweight, and (c) ultralight wheelchairs.

pain, or the progression of a disease such as Multiple Sclerosis or Muscular Dystrophy. Recently, the development of Pushrim Activated Power Assist Wheelchairs (PAPAWs) has provided an alternative for these users. The PAPAW works through the use of a battery and a motor mounted directly into the wheel. The motor provides a supplement to the user so that very little force input from the user will still afford normal momentum. Transitioning from a manual wheelchair to a powered wheelchair may be difficult both physically and psychologically for some people. Users may not want to modify their homes or their cars to accommodate a powered wheelchair and may be used to providing their own mobility. Although the PAPAW provides a good intermediate wheelchair for users who may still benefit from a manual wheelchair, but do not have the strength or stamina, it also has its disadvantages. The added weight of the motor driven wheels dramatically increases the overall weight of the wheelchair and can also be difficult for users during transfers. Additionally, algorithms for the control of the PAPAWs are not yet refined and can lead to difficulty propelling the wheelchair.

POWERED WHEELCHAIRS

Powered wheelchairs represent devices that can provide an incredible amount of independence for individuals who may have extremely limited physical function. Perhaps even more than manual wheelchairs, having the proper elements and setup of a power wheelchair is vital. The lifestyle of the user, the activities in which they would like to participate, the environments to which they will be subjected, and ability level have all contribute to the correct prescription of a powered wheelchair. Many adjustments can be made to any particular powered wheelchair to specifically fit the needs of the user. Seating systems, cushions, added assistive technology to name a few. This section will focus on the characteristics that differentiate certain powered wheelchairs from one another (Fig. 2).

Powered wheelchairs come in three drive wheel setups: front, mid, and rear wheel. Each of these setups has different advantages and shortcomings. Powered wheelchair users are each unique and have specific requirements for their activities of daily living, such as maneuverability, obstacle climbing, or driving long distances. Mid-wheel drive wheelchairs tend to provide greater maneuverability because the drive wheels are located directly beneath the user's center of mass. Front-wheel drive wheelchairs are often associated with greater obstacle climbing ability. Rear-wheel drive powered wheelchairs are good for speed and outdoor travel, but may not provide the maneuverability or stability for some users. The different wheelchair setups provide the means for users to achieve their goals.

For the user, the joystick is probably the most important part of the powered wheelchair. However the standard displacement joystick is not acceptable for all users. Current technologies have allowed almost any user to operate a powered wheelchair as well as other assistive technologies, such as computers. Even the smallest abilities of the user can be translated to powered wheelchair mobility such as foot control, head control, finger control, tongue control and so on.

Like manual wheelchairs, powered wheelchairs also have different classifications. Medicare defines powered wheelchairs in three major categories: Standard weight frame powered wheelchair (K0010), Standard weight framed powered wheelchair with programmable control parameters for speed adjustment, tremor dampening, acceleration control and braking (K0011), other motorized powered wheelchair base (K0014) (17). Pearlman et al. (18) recently conducted a study examining the reliability and safety of standard weight frame powered wheelchairs. Of the 12 wheelchairs tested, only 3 passed the Impact and Fatigue section of the ANSI/RESNA Standards (19,20). Medicare has recently proposed to stop funding these powered wheelchairs, recommending the addition of a programmable controller that would place it in the second category (K0011). However, this may not be acceptable since the problems exist mainly with the drive train or the frame of the wheelchair and these parameters would not change. In order to adequately serve the interests of powered wheelchair users, the frames, drive trains, and control systems all need to perform within the standards put forward by ANSI/RESNA.

WALKERS AND ROLLATORS

Some individuals may have the ability to ambulate, however, they may get tired very easily or have visual

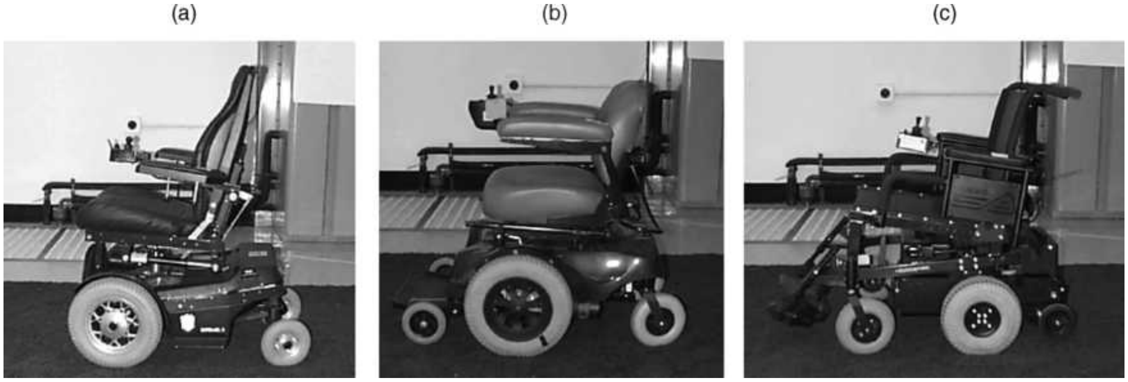


Figure 2. (a) Front-, (b) mid-, and (c) rear-wheel drive powered wheelchairs.

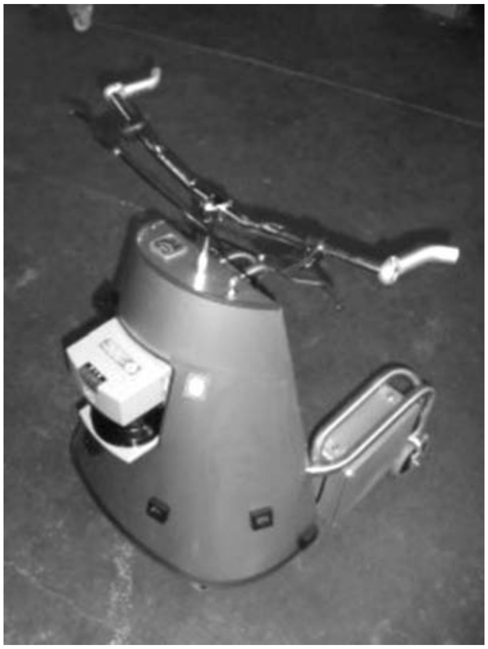


Figure 3. Robotic walker for users with visual impairments.

impairment problems that may result in a fall and injury. Fuller (21) reported that 33% of community-dwelling elderly people and 60% of nursing home residents fall each year. Walkers and rollators represent useful assistive technology devices for these individuals by lending support and weight relief during mobility. They may have zero, two, or four wheels and possibly have hand brakes. They may also contain a small area for sitting if the user becomes fatigued or a basket for carrying items for the user.

Some elderly persons, in addition to having mobility impairment, also have a visual impairment. Recent research has investigated a new robotic walker (Fig. 3) that through the use of sonar and infrared (IR) sensors can detect obstacles as well as provide guidance along a preprogrammed path (i.e., maneuvering around an assisted living home) (22).

SPORTS AND RECREATION DEVICES

As quality of life receives more attention, sports and recreational activities have become more important to individuals with disabilities. Sport and recreational activity participation provides many benefits to individuals with disabilities. Physical activity reduces or slows down the development of cardiovascular disease as well as modifies risk factors including high blood pressure, blood lipid levels, insulin resistance, and obesity (23). In addition, the development of muscular strength and joint flexibility gained through regular exercise improves the ability to perform activities of daily living (24). Regular exercise may help reduce clinical depression and days spent as an in-patient in a hospital, and may improve social interactions and prolong life expectancy (25). With the positive benefits of sports, exercise, and recreational activities in mind, the purpose of this section is to describe some of the more popular sports played by individual with disabilities.

Wheelchair Basketball

Wheelchair users who play basketball may have various diagnoses, such as paraplegia, cerebral palsy, amputations, post-polio syndrome, or a disabling injury. Participants are not required to use a wheelchair for their primary means of mobility or in their activities of daily living. Prior to the actual game, persons who want to play basketball must have their player classification level determined by a qualified referee. To equalize the capability of each team, the classification levels of the competitors are matched (26).

Whether players play zone or person-to-person basketball, the basic rules apply to both. Because different players have varying degrees of disability, rules have been developed that all players need to abide by. Keep firmly seated in the wheelchair at all times. A player may not use a functional leg or leg stump for physical advantage. An infraction of this rule constitutes a physical advantage foul (27).

Wheelchair basketball is similar to an everyday wheelchair, but incorporates features that enhance maneuverability (Fig. 4). Basketball wheelchairs are lightweight to allow for speed, acceleration and quick braking. The wheelchair must have four wheels. Two large, rear wheels and two front casters. The front casters are 2 in. (50 mm) in diameter and typically made from hard plastics, similar to the material used to make inline skate wheels. The rear wheels must be larger than or equal to 26 in. (338 mm) in diameter. The rear wheels must have handrims. Basketball wheelchairs use spoke guards made of high impact plastic. These guards cover the rear wheel spokes to prevent wheel damage and illegal ramming and picking. The spoke guards provide several benefits: First, spoke guards can be used to pick up the ball from the floor. Using a hand, the player pushes the ball against the spoke guard and rolls it onto their lap, Second, spoke guards protect hands and fingers from injury when reaching for the ball near another player's rear wheel. Third, they provide space to identify team affiliations and sponsor names. Camber is an important feature of basketball wheelchair as well. Camber is defined as"the angle of the main wheel to the vertical", or as a situation in which"the spacing between the top

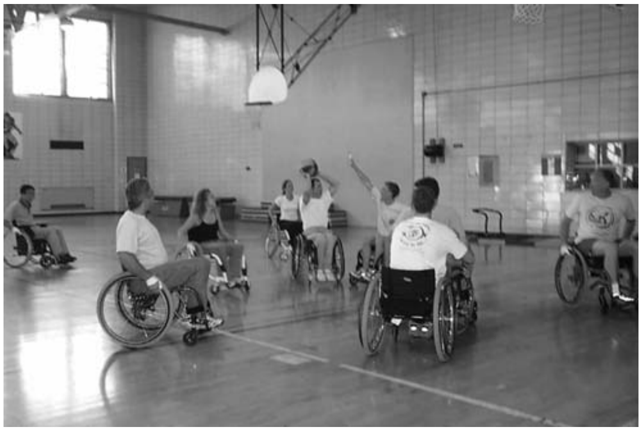


Figure 4. Players competing in a friendly game of wheelchair basketball.

points of the wheels may be less than the spacing between the bottom points". Increasing camber slightly reduces the height of the seat, while it proportionally increases the wheelbase, which corresponds to the width of the wheelchair. In the same way, with negative camber, the center of gravity of the occupied wheelchair moves backward. From a practical point of view, increased wheel camber improves hand protection as chairs pass through doors and, in terms of basketball, camber makes a wheelchair more responsive during turns and protects players' hands when two wheelchairs collide from the sides, by limiting the collision to the bottom of the wheels and leaving a space at the top to protect the hands. Basketball wheelchair seats typically have a backward seat angle slope of 5–15°. The angle of the seat compared to the ground is known as "seat angles". Guards are an exception. Guards are allowed to have lower seat heights and greater seat angles. These modifications make chairs faster and more maneuverable for ball handling.

Wheelchair Racing

Individuals with all levels of SCI as well as lower limb amputees can participate in competitive races. The preferred racing chair among racers is the three-wheel chair (Fig. 5). The three-wheel design is constructed from high pressure tubular tires, light weight rims, precision hubs, carbon disk/spokes wheels, compensator steering, small push rings, ridged aluminum frame, and 2–15° of wheel camber. The camber in a racing chair makes the chair more stable and allows the athlete to reach the bottom of the pushrim without hitting the top of the wheels or pushrim.

Hand-Cycling

Cycling has been a popular outdoor sport for several years. The adaptability of cycling to different terrains makes it a favorite for many. Adaptive equipment for bicycles consists of a hand cycle allows individuals with limited use of their legs to utilize the strength of their arms (28). A handcycle typically consists of a three-wheel setup to compromise for the balance required when riding a two-wheeled bicycle. Two-wheeled handcycles do exist but require a great deal of skill and balance. Handcycle designs allow the user to propel, steer, break, and change gears, all with the upper extremities and trunk. Two types of handcycle designs are readily available (1) the upright and (2) the recumbent. In an upright handcycle, the rider remains in an upright position similar to the position the body takes when seated in a touring bike. Upright handcycles use a pivot steer to turn. Only the front wheel turns while the cycle remains in an upright position. Transferring and balancing tend to be easier on the upright cycle. In the recumbent handcycle, the rider's torso reclines and the legs are positioned out in front of the cyclist. These cycles use a lean-to-steer mechanism. The rider leans to turn, causing the cycle to pivot at hinge points. Leaning to turn can be challenging if the rider lacks trunk stability, in which case a pivot steering recumbent handcycle may be more appropriate. Recumbent handcycles are lighter and faster, making them the choice for hand cycle racing. Relatively minimal modifications are needed to accommodate individuals with tetraplegia. Some of the modifications include hand cuffs that can be mounted to the arm crank handles and elastic abdominal binders which can be fitted around the user and the handcycle seat to increase trunk stability.

Wheelchair Rugby

Rugby is played indoors on a large gym on a basketball court surface. Players use manual wheelchairs specifically designed for the sport. Due to the level of contact, the chairs have protective side bars on them and players are strapped in to prevent injury. Most chairs are made of titanium or steel to handle the hits that they sustain. In addition, the low pointers have a high camber (angle of the wheels) so that they can turn fast, as well as "red push rim covers so they can actually stick to the other person's chair and hold them." The high pointers have armor on the front of their chairs resembling a cow catcher so that they can push through the other players without getting stuck (Fig. 6).

To be eligible to play rugby, players must have a combination of upper and lower extremity impairment. Most of the players have sustained cervical level spinal injuries and have some degree of tetraplegia. Like in basketball, players receive a classification number based on there level of impairment (29). Rugby consists of two teams comprised of four players. The object of the game is for a player to have possession of the ball and cross the opponent's goal line.

Rugby wheelchairs are strictly regulated to ensure fairness. However, chairs may vary considerably depending on

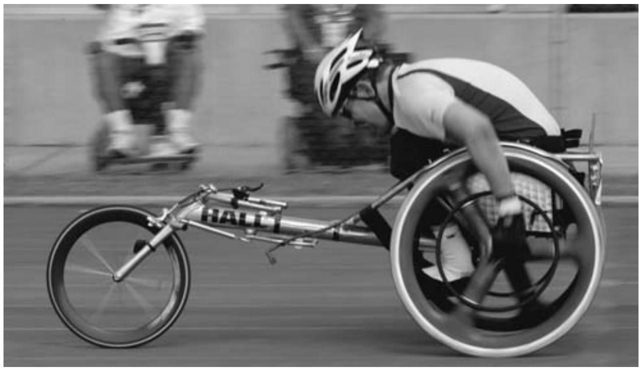


Figure 5. Shows a racing three-wheeled wheelchair.

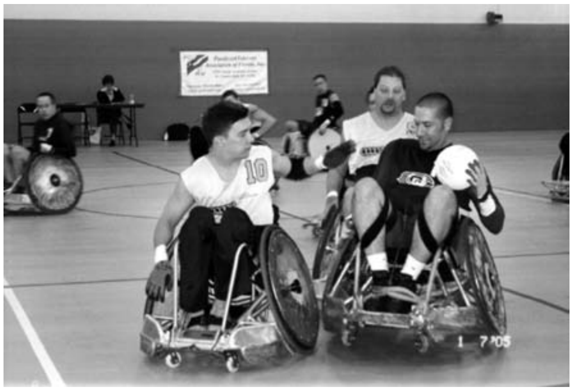


Figure 6. Wheelchair rugby.

a player's preferences, functional level and team role. Team roles may be assigned according to ability. Players with upper body limitations tend to perform the defensive blocking and picking roles. They use chairs that have additional length and hardware. All rugby chairs have extreme amounts of camber, 16–20°, significant bucketing, and antitip bars. The camber provides lateral stability, hand protection, and ease in turning. The bucketing (knees held high relative to rear end) helps with trunk balance and protection of the ball.

Tennis

Tennis players compete in both singles and doubles games. Players are required to have a permanent mobility-related physical disability that requires a wheelchair as the primary means of mobility. Tennis is played on the traditional tennis court using the tradition size and height tennis net. However, unlike traditional tennis, the ball is permitted two bounces on the court before it must be returned. Brakes are not permissible as stabilizers and the athlete must keep one buttock in contact with the seat at all times.

Tennis players use a three-wheeled chair with a large amount of camber to maximize mobility around the court. The seat is situated at a steep backwards seat angle slope. The angle helps with balance, keeps players against the seat backs, and gives them greater control over the wheelchair. The knees tend to be flexed with the feet on the footrest behind the player's knees. With the body in a relatively compact position, the combined inertia of rider and wheelchair is reduced, making the chair more maneuverable (30). Handles and straps can also be added to the chair. Many players incorporate plastic rigid handles into the front of the seat. Players use these handles when leaning for a shot or making quick directional changes. Straps can be used around the waist, knees and ankles, to help with balance (31).

Adaptive Skiing

Skis for skiers with disabilities have advanced state-of-the-art skis that offer shock absorption systems, frames molded to body shape, and quick release safety options. Skiers with disabilities can maintain a similar pace to that of unimpaired athletes with the development of adaptive seating, backrests, cushions, tethering ropes, roll bars and outriggers. Outriggers, an adapted version of a forearm crutch with a shortened ski, provide extra balance and steering maneuverability (32). Two types of sit-down adaptive skies are available: Bi and Mono. Bi skis are appropriate for skiers with limited trunk stability. With Bi skis, the skier balances on two skies and angulates and shifts to put the skis on edge. Bi skis have wider base of support, can usually be mastered quickly with few falls and are easier to control than a mono ski. The Mono Ski is the ski of choice for individuals who want high end performance, maneuverability and speed. With a mono ski, the skier sits relatively high on the seat of the ski over the snow. The skier uses upper body, arm and head to guide their movement down the hill. Sit-, Mono- and Bi-skis have loading mechanisms, usually hydraulic, that enable the individual to raise themselves to a higher position for transferring onto a ski lift.

TRANSPORTATION SAFETY AND ADAPTIVE DRIVING FOR WHEELCHAIR USERS

Wheelchair users, like the entire population, use several forms of transportation to travel from place to place: They are passengers in public transportation systems (bus, subways, and vans) and private vehicles, and are potential drivers of each of these types of vehicles. To ensure the safety of *all* passengers and drivers, certain safety mechanisms must be in place: drivers must be able to safely control the vehicle, and all seated passengers require seats securely fastened to the vehicle and passenger restraints (e.g., seatbelts) which can secure the passenger to the seat. The requirement of passenger restraints is relaxed for passengers in large vehicles, like busses and subways, because of the low likelihood of high velocity crashes (33). In many cases, adaptation of a vehicle is necessary when the original equipment manufacturer (OEM) control and/or securement mechanism cannot provide adequate safety for wheelchair users because either (1) sensory and/or motor impairments of the user requires adaptive driving equipment so the vehicle can be safely controlled, or (2) the user cannot safely or effectively use the OEM seat or passenger restraint system in the vehicle.

Vehicle Control Systems

The complexity of the adaptive equipment required for safe control of the vehicle is correlated to the type and level of impairment of the wheelchair user. Adaptive driving equipment can be as low tech as attaching a knob on a steering wheel, and as high tech as fly-by-wire technology, where computer-controlled actuators are added to all of the controls, and the drive interfaces with the computer (via voice and/or low or no-effort sensors). Adding actuators to all of the driving and operating controls of the vehicle through computer controls.

An example of this range is the types of steering adaptations available for people with disabilities. Figure 7 demonstrates both a typical steering knob for a person with little to no loss of hand sensory-motor function (a), and (b) a knob for someone with loss of some sensory motor function. Both types of steering knobs serve the same purpose: They allow the driver to safely steer the vehicle with one hand while (typically) their other hand is operating a hand control which actuates the fuelaccelerator and brake pedals. When upper-extremity sensory-motor function does not allow for safe turning of the OEM steering system (even with a knob), actuators can be used in lieu of upper-extremity function. These systems are named "low-" or "no-effort" steering systems, depending on the type of assistance that the actuators provide. Retrofitted controls for these systems are used and typically require removal of the OEM equipment (e.g., the steering wheel and/or column). Consequently, when this level of technology is used, it usually becomes unsafe for an unimpaired individual to drive the vehicle (without significant training).

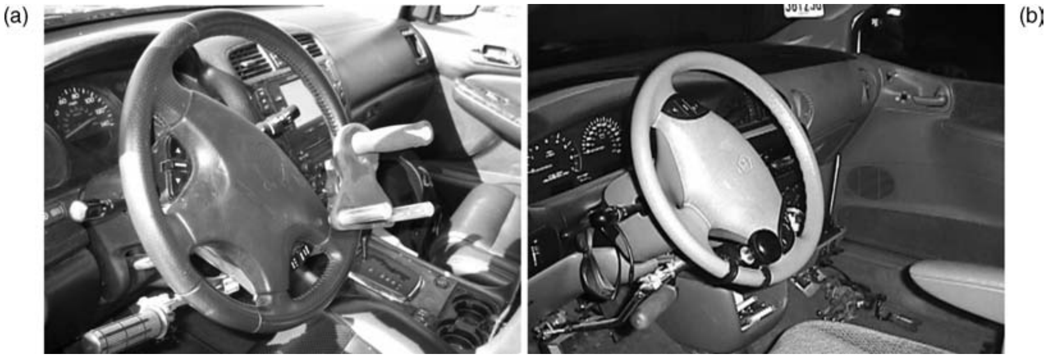


Figure 7. Steering knobs.

Common fuelaccelerator and braking system hand controls are bolted to the OEM steering column, and actuate each pedal with mechanical rods (Fig. 8). These types of controls require nearly complete upper extremity function to operate. When upper extremity function is substantially impaired actuators are added to the braking and fuelaccelerator systems and are operated through some switching methods. The types of switches depend on the most viable control mechanism for the user: in some cases, a simple hand-operated rheostat variable resistance switch (i.e., dimmer switch or rheostat) may be used, and in other cases, a breath-activated pressure switch (sip-and-puff) system may be used.

The above steering, fuelaccelerator, and brake adaptive equipment are focused on the primary control systems of the vehicle (those which are required to drive the automobile). Various adaptive equipment can control the secondary control system of the vehicle also (e.g., ignition switch, climate control, windows). Like the primary control adaptive equipment, equipment to modify the secondary controls of the vehicle range from low to high tech. For example, users with impaired hand function may require additional hardware to be bolted to the ignition key so they can insert the key and turn it to start the vehicle. Alternatively, hardware can be added to allow a drive to start the vehicle via a switch, either remotely or from within the cabin. Power windows and door-lock switches can be rewired to larger switches, or in more accessible locations for the driver.

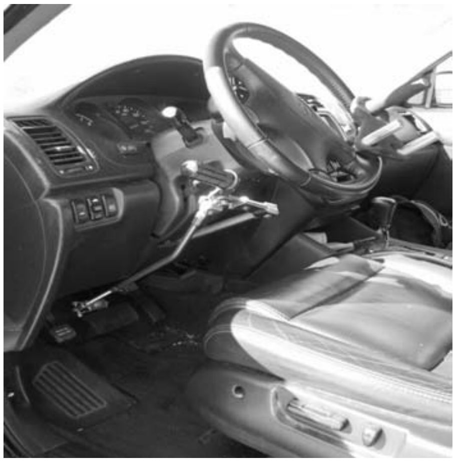


Figure 8. Common hand controls.

Both primary and secondary control systems must be placed in locations easily accessible to the driver. In some cases, wheelchair riders will transfer out of their wheelchair directly into the OEM seating system, allowing for most controls to remain in their OEM locations. If a wheelchair user remains in their wheelchair and drives the vehicle the controls must be made accessible to their seated position and posture. In these cases, along with the case a wheelchair users remaining in their wheelchair as passenger, provisions must be made to safely secure the wheelchair to the vehicle, and to provide adequate passenger restraints.

Wheelchair Tie-Down and Occupant Restraint Systems (WTORS)

For both practical and safety reasons, when a wheelchair user remains in wheelchair while riding in a vehicle they must be safely secured to the vehicle. An unsecured wheelchair will move around, causing the user to be unstable while the vehicle is moving. Being that power wheelchairs can be in excess of 200 lb (91 kg) in weight, This instability could cause harm to the wheelchair rider and/or the surrounding other passengers vehicle occupants. If the wheelchair user is driving, they may lose control of the vehicle if they accidentally roll away from the vehicle controls. Another important concern for an unsecured wheelchair rider is in the case of an accident. An unsecured wheelchair and user can easily be ejected out of the vehicle if they are not secured. The WTORS systems are currently governed by the ISO 7176-19 (34).

Several types of tie-down systems exist, including a four-point belt system and various latching-type mechanisms which typically require hardware to be attached to the wheelchair. The four-point belt systems are most common, and are found on public busses, and also private vehicles. Theses systems are the most widely used tie-down system because they can be attached to a wide variety of wheelchairs. In some cases, manufacturers incorporate attachment rings for these tie-down systems into their wheelchair. When no points of attachment are available (most common situation), points at the front and rear of the seat or frame be used. These attachment points must be sufficiently strong to secure the wheelchair in the event of a crash, and be in locations which will allow the straps to be oriented within a specified range of angles with the horizontal (front straps: 30–60°, rear: 30–45°).

Unfortunately, tie-down systems are they are not convenient to use: a second person (other than the wheelchair user) is typically needed to help secure the wheelchair, making the operation laborious, and in some cases awkward for the wheelchair users who may not be comfortable with another person touching their wheelchair or encroaching on their personal space. Consequently, and especially on public transportation, these systems are commonly unused and the wheelchair user relies on their brakes wheel-locks for stability, risking their own safety in a crash.

Other mechanisms have been used to secure a wheelchair to the vehicle. These included wheel-clamps and t-bar systems. With these mechanisms, wheelchairs are secured to the vehicle through a mechanical clamp that adjusts to the wheelchair size. These systems are quicker to attach to the wheelchair, but are still difficult or impossible for a user to use independently.

A variety of wheelchair tie-down systems have been developed that allow the wheelchair users to independently lock and unlock their wheelchair to the vehicle. A common one used for personal vehicles is the EZ-Lock System, which is a hitch system for the wheelchair. This system allows the wheelchair user to maneuver the wheelchair so a specialized hitch attached to the wheelchair is captured into a latch bolted to the vehicle; both electric and manual release mechanisms can be used to unhitch the wheelchair from the device, allowing for custom placement of a hitch for easy accessibility to the wheelchair user. A drawback to this system is the specialized hardware that must be attached to the wheelchair that restricts folding a manual wheelchair and reduces ground clearance.

Because this system is designed to allow the user to drive forward into the device, it works well and is common in private vehicles where the wheelchair user drives the vehicle. In larger vehicles, such as public busses, it is typically more convenient for a user to back into a spot and lock their wheelchair.

An ideal system for public transportation would be one that a user can operate independently and that does not require specific hardware to be attached to the wheelchair that may not work on *all* wheelchair models. A system is currently being developed at the University of Pittsburgh that tries to achieve these goals.

Occupant restraint systems are the last requirement to allow a wheelchair user to safely travel in a vehicle. These restraint systems mimic the function of a seat belt, and can be either attached to the wheelchair (integrated restraint) or to the vehicle (Fig. 4). In both cases, the placement of the restraints with respect to the body is critical to prevent injury in a crash—either through direct insult of the seatbelt with the body, or because of submarining (where the torso slides down under the pelvic belt).

To ensure these WTORS systems and the wheelchair themselves can safely survive a crash, standards testing is in place. Rehabilitation Engineering and Assistive Technology Society of North America (RESNA), International Standards Organization (ISO), and Society of Automotive Engineers (SAE) have worked in parallel to establish minimum standards and testing methods to evaluate wheelchairs and WTORS systems (35). These tests mimic those performed on OEM seat and occupant restraint systems, which suggest the system should be able to withstand a 20 g crash (36). To encourage and guide wheelchair manufacturers to build their wheelchairs to these standards researchers have developed a website to inform all relevant stakeholders of the latest information (http://www.rercwts.pitt.edu/WC19.html).

LOWER EXTREMITY PROSTHETICS

Prosthetics are devices that replace the function of a body organ or extremity, unlike orthotic devices, which support existing extremities. Prosthetics range from simple cosmetic replacements to complicated structures that contain microprocessors for controlling hydraulic and pneumatic components. Commonly used prosthetic devices primarily include artificial limbs, joint implants, and intraocular lenses. Approximately, 29.6–35.4% of the U.S. population use prosthetic limbs (37) with >2% of them aged between 45 and 64 years using lower extremity (LE) prosthetics for mobility (U.S. Bureau of Census, 2000). Amputation, resulting from peripheral vascular diseases in the older population (60 years and older) and trauma in young population can be considered factors for the use of LE prosthetics.

Research and development in clinical practice has resulted in recent advances in the area of prosthetics designs and controls technology. Examples of these advances include the use of injection molding technology for socket manufacturing, shock absorbing pylons, the incorporation of neuro-fuzzy logic microprocessor-based controllers for myoelectric prostheses, and microprocessor-controlled prosthetic knees and ankles (38,39).

Prosthetic feet classified as"uniaxial" allow for movement at a single axis in one plane, such as plantarflexion and dorsiflexion of the ankle. In this type of prosthetic foot, the heel is typically composed of the same density materials as the rest of the foot, with an option of different heel height. Also, uniaxial feet have different options at the rubber toe section in terms of flexibility, which depends on the weight of the individual. Multiaxial prosthetic feet (MPFs) have five degrees of motion in three planes: plantarflexion–dorsiflexion, inversion/eversion, and rotation. This feature provides stability to the user, while walking on uneven surfaces and also aid in shock absorption lowering intensity of shear forces on residual limb. Elastomer or rubberized material is used to alter, resist, or assist with the different degrees of motion in the prosthetic foot. MPFs also provide options for different heel heights [0.5–1 in. (13–26 mm)] and different degrees of toe material resistance, while split internal structures in the heel assist with inversion/eversion on uneven ground. The MPFs are prescribed by the weight and shoe size of the consumer.

The solid ankle, cushion heel (SACH) prosthetic foot is the most commonly prescribed prosthetic foot for lower extremity amputations. The SACH foot is constructed out of eurothene (plastic) materials with a less dense material

Table 1. **Functional Level and Devices**

	Functional Level	Type of Device
K0	No ability to ambulate or transfer safely; prosthesis does not enhance mobility	Cosmesis
K1	Transfers and ambulates on level surfaces; household use	SACH
K2	Able to negotiate over low level environmental barriers; limited community ambulation	Low level energy storage feet
K3	Prosthetic usages are beyond simple ambulation; able to traverse MOST environmental barriers and is a community ambulator	Energy storage prosthesis
K4	Able to perform prosthetic ambulation exceeding basic skills (i.e., high impact); child, active adult and athlete	Energy storage prosthesis

incorporated at the heel. Use of materials with different densities permits proper positioning while standing, as softer heels aides in enhancement of the walking efficiency after heel strike, by shifting center of gravity forward. A device known as durometer is used to measure the density of plastics used in prosthetic devices. The weight and activity level of the individual using the prosthesis determines which heel density is selected, as heavier user require firmer heel cushion. The stiffness of heel, also determine amount of knee flexion and shock absorption. Greater the heel stiffness more the knee flexion and lower shock absorption during heel strike and vice versa. The SACH foot also contains a keel made out of a hard wood or composite material. Belting material is applied to the keel, which prevent the keel it from breaking through the eurothene cover. During ambulation, the foot simulates plantar flexion movement and prevents the loss of anterior support during the push off at the toe.

Individuals who use foot prosthetic devices are assessed for weight, potential activity levels, and type of use for which they anticipate using their prosthetic devices. Based on this assessment, clients are then categorized into four functional levels:

Energy Storage and Return (ESAR) prosthetic feet are fabricated to assist with the dynamic response of feet, acting as a diving board from which a person can push off during walking. These feet have capability to store energy during stance phase and return it to the user to assist in forward propulsion in late stance phase. The ESAR has flexible keels and are prescribed by the anticipated activity level and weight of the person. Also, limited evidence suggests use of ESAR as their use results in increasing ambulation speed and stride length ~7–13% greater than with a conventional (SACH) foot in both traumatic and vascular transtibial amputees (40).

Macfarlane et al. (40) compared energy expenditure of individuals with transfemoral amputations who walked with a SACH foot versus a Flex-Foot prosthetic. The SACH has a solid ankle and cushioned heel construction, while the Flex-Foot prosthetic has a hydraulic knee joint. The authors determined that Flex-Foot walking resulted in significantly lower exercise intensity, reduced energy expenditure and improved gait efficiency. These findings are significant considering the SACH foot is the most commonly used foot prosthetic in the U.S. today (41). Lower energy expenditure was also reported for individuals with trans-tibial amputation with the use of Flex-Foot as compared with a SCAH foot.

Higher level of limb loss results in addition of more prosthetic components. Prostheses for transfemoral amputations comprised of four basic components: the socket, the knee joint, the pylon, and the foot. Pylons are classified as: exoskeleton in which the weight of the individual is supported by the external structure of the prostheses (i.e., a crustacean shank), or endoskeleton that is comprised of an internal, weight-bearing pylon encased in moldable or soft plastics (i.e., modular pylon). The knee mechanism use, a conventional damping system, where a flow of (fluid or air) is controlled by a valve and its operation is set for a particular walking speed according to user's preference. The system described as intelligent prosthesis (IP), where a diameter of damping controlling valve is changeable according to varying speed of walking (42). Romo provided guidance on the selection of prosthetic knee joints and indicated that proper alignment impacts the effectiveness of matched and adjusted knee joints for smooth and reliable gait (43). Taylor et al. (44) compared effectiveness of an intelligent prosthesis (IP), and pneumatic swing-phase, control-dampening systems while walking on a treadmill at three speeds of 1.25, 1.6, and 2 mph (2, 2.6, and 3.2 $\text{km} \cdot \text{h}^{-1}$). The results indicated lower VO_2 consumption for individuals using IP compared to controls-damping system at 2 mph (3.2 $\text{km} \cdot \text{h}^{-1}$). The question often raised by critiques concerns the cognitive demands by the high end technology on the users. The results of the study by Heller et al. (42) that investigated cognitive demand when using the IP compared to a conventional prosthesis indicated no significant differences while using these prostheses for ambulation. Though not uncommonly prescribed high rejection rates has been described for prostheses after hip disarticulation and hemipelvectomy. These prostheses consist of an addition of hip joint mechanism to other parts similar to prostheses prescribe after transfemoral amputation.

Modular systems were first developed in the 1960s by Otto Bock, which consisted of shock absorbing pylons that contained with shock absorbers. Also, a reverse-pyramid fixture at the ends of the pylon permits angular adjustments to the alignment of these devices with the residual limb. Modular systems are lighter than the earlier wooden systems, allow for 15° of movement gain in either the frontal or sagittal plane, and also permit internal and external rotational adjustments. Modular systems can extend the life of a prosthetic device, as worn parts can be replaced. In addition, individuals using these systems experienced less need for maintenance.

A significant improvement in the design procedure of the prosthetics considers the interaction of forces between prosthesis and residual limb can be found in the literature. Jia et al. (45) studied the exchange of loads and forces between the residual limb and prosthetic socket in transtibial amputation using the Finite Element Analysis (FEA) method. Lee et al. (46) used FEA to determine contact interface between the transtibial residual limb and prosthetic socket. The study determined the need for sameness of shapes for both the residual limb and socket in order to decrease peak normal and shear stresses over the patellar tendon, anterolateral and anteromedial tibia, and popliteal fossa. Winson et al. investigated the interaction between socket and residual limb during walking using a FEA model for transtibial prosthesis. Pylon deformities and stress distribution over the shank were problems identified during walking and results indicated need for pylon flexibility for better optimization and need of future studies identifying fatigue life of these prostheses (47).

With advancement in the area of prosthetics designs and development, simultaneous factors that need to be considered, use of these devices in clinical practice for targeted population and cost containment. Premature abandonment of mobility assistive devices, which might be due to poor performance and/or changes in the need of the user, is not uncommon and adds to the expense of these devices (48). Improved quality of service delivery for LE prostheses, which include identifications of reasons for successful use or nonuse of LE prostheses, is needed (49). Also, incorporation of standardized performance testing procedure to ensure durability of LE prosthetics is vital to the appropriate prescription of, and satisfaction with, prosthetic devices.

Prosthetic devices of today incorporate advancements from the aerospace and engineering fields and include the use of new materials, such silicone elastomer gel sleeves in to assist in the fit of prosthetic sockets, prosthetic feet made from carbon-fiber composite components that are lighter in weight, and surgical implantation of titanium prosthetic attachment devices directly to bones of residual limbs (50,51). Neuro- and microprocessors and sensors are now incorporated on-board the prosthetic device to control knee joint movement to improve the symmetry of different gait patterns across a variety of cadence speeds. Hydraulic or pneumatic devices are also used to dampen the swing-through phase of walking with the prostheses to assist with walking at difference cadences (52,53). Manufacturers are now using computer-aided design and manufacturing techniques to improve the fit of the prosthetic sockets as well as component designs (54,55).

Because of the growing population of people in need of mobility aids, and their demand to maintain their lifestyle, whether that includes going to and from work, participating in extracurricular activities, or maneuvering around their environment, continuing information must be gathered and disseminated to make these goals achievable. Through technological advancements people who require mobility aids can accomplish more of their goals than ever before, however there are still people for whom the technology is not yet developed enough or cannot obtain the proper devices to meet their needs. It is for this reason that problems must continually be studied and innovations must advance so that mobility aids will serve anyone who requires them to meet their goals.

BIBLIOGRAPHY

1. Schiller JS, Bernadel L. Summary Health Statistics for the U.S. Population: National Health Interview Survey, National Center for Health Statistics. Vital Health Stat 10(220):2004.
2. Sie IH, Waters RL, Adkins RH, Gellman H. Upper extremity pain in the postrehabilitation spinal cord injured client. Arch Phys Med Rehab 1992;73:44–48.
3. VanSickle DP, Cooper RA, Boninger ML, DiGiovine CP. Analysis of vibrations induced during wheelchair propulsion. J Rehab R&D 2001,38:409–421.
4. Calder CJ, Kirby RL. Fatal wheelchair-related accidents in the United States. Am J Phys Med Rehab 1990;69:184–190.
5. Mills T, et al. Development and consumer validation of the Functional Evaluation in a Wheelchair (FEW) instrument. Disabil Rehab 2002;24(1–3):38–46.
6. Chase J, Bailey DM. Evaluating potential for powered mobility. Am J Occup Therapy 1990;44(12):1125–1129.
7. Cooper RA, Cooper R. Electric Powered Wheelchairs on the Move. Physical Therapy Products; July/August, 1998, p 22–24.
8. Galvin JC, Scherer MJ. Evaluating, Selecting, and Using Appropriate Assistive Technology. Gaitherburg (MD):Aspen Publishers Inc.; 1996.
9. Cooper RA. Rehabilitation Engineering Applied to Mobility and Manipulation. Bristol (UK): Institute of Physics; 1995.
10. Grieco A. Sitting posture: An old problem and a new one. Ergonomics 1986;29:345–362.
11. Cooper RA. A perspective on the ultralight wheelchair revolution. Tech Disab 1996;5:383–392.
12. Cooper RA, et al. Performance of selected lightweight wheelchairs on ANSI/RESNA tests. Arch Phys Med Rehabil 1997;78:1138–1144.
13. Masse LC, Lamontagne M, O'Riain. Biomechanical analysis of wheelchair propulsion for various seating postitions. J Rehab R&D 1992;29:12–28.
14. International Standards Organization, Evaluation of Human Exposure to Whole-Body Vibration—Part 1: General Requirements. ISO 2631-1, Washington (DC): ANSI Press; 1997.
15. Wolf EJ, et al. Analysis of whole-body vibrations on manual wheelchairs using a Hybrid III test dummy. Proceedings of the annual RESNA conference; 2001. p 346–348.
16. Kwarciak A, Cooper RA, Wolf EJ. Effectiveness of rear suspension in reducing shock exposure to manual wheelchair users during curb descents. Proceedings of the annual RESNA conference; 2002. p. 365–367.
17. Centers for Medicare and Medicade Services (CMS), http://www.cms.hhs.gov/providers/pufdownload/anhcpcdl.asp, Accessed 2005. Jan 12.
18. Pearlman J, et al. Economical (K0010) Power Wheelchairs Have Poor Reliability and Important Safety Problems: An ANSI/RESNA Wheelchair Standards Comparison Study. Proceedings of the Annual RESNA Conference; 2005.
19. American National Standard for Wheelchairs, Volume 2, Additional Requirements for Wheelchairs (Including Scooters) With Electrical Systems. Virginia: Rehabilitation Engineering and Assistive Technology Society of North America; 1998.

20. American National Standard for Wheelchairs, Volume 1, Requirements and Test Methods for Wheelchairs (Including Scooters). Virginia: Rehabilitation Engineering and Assistive Technology Society of North America; 1998.
21. Fuller GF. Falls in the Elderly Am Fam Physician 2000;61(7):2159–2168.
22. Rentschler AJ, et al. Intelligent walkers for the elderly: Performance and safety testing of the VA-PAMAID robotic walker. J Rehab R&D 2003;40(5):423–432.
23. Abel T, et al. Energy Expenditure in wheelchair racing and hand biking- a basis for prevention of cardiovascular diseases in those with disabilities. Eur J Cardio Prev Rehab 2003;10(5):371–376.
24. Franklin BA, Bonsheim K, Gordon S. Resistance training in cardiac rehabilitation. J Cardiopulm Rehab 1991;11:99–107.
25. Kenedy DW, Smith RW. A comparison of past and future leisure activity participation between spinal cord injured and non-disabled persons. Paraplegia 1990;28:130–136.
26. National wheelchair basketball association (2003–2004) official rules and case book retrieved from National Wheelchair Basketball Association. http://www.mwba.org. Accessed.
27. Vanlandewijck Y, Daily C, Theisen DM. Field test evaluation of aerobic, anaerobic, and wheelchair basketball skill performances. Int J Sports Med 1999;20(8):548–554.
28. Janssen TWJ, Dallmeijer AJ, Van der Woude LHC. Physical capacity and race performance of handcycle users. Jour Rehab R&D 2001;38(1):33–40.
29. Lapolla T. International rules for the sport of wheelchair rugby. http://quadrugby.com/rules.htm. Accessed 2000.
30. Wheelchair tennis handbook. International Tennis Federation. http://www.itfwheelchairtennis.com. Accessed 2004.
31. Kurtz M. Difference makers. Sports' N Spokes 2002;28(2):10–14.
32. Russell JN, et al. Trends and Differential use of Assistive Technology Devices: United States, 1994. Adv Data 1997; 292:1–9.
33. Shaw G, Gillispie B. Appropriate portection for wheelchair riders on public transit buses. J Rehab R&D 2003;40(4):309–320.
34. International Standards Organization, Wheeled mobility devices for use in motor vehicles. ISO 7176-19, Vol. 31. 2004.
35. Hobson D. Wheelchair transport safety - the evolving solutions. J Rehab R&D 2000;37(5).
36. Bertocci G, Manary M, Ha D. Wheelchair used as motor vehicle seats: Seat loading in frontal impact sled testing. Med Eng & Phys 2001;23:679–685.
37. Mak AF, Zhang M, Boone DA. State-of-the-art research in lower-limb prosthetic Biomechanics socket interface: a review. J Rehab R&D 2001;38(2):161–174.
38. Weir RF, Childress DS, Grahn EC. Development of Externally-Powered Prostheses for Persons with Partial Hand Amputations. Proceedings of the Chicago 2000 World Congress on Medical Physics and Biomedical Engineering; 2000 July 23rd–28, Chicago (IL):.
39. van der Linde H. A systematic literature review of the effect of different prosthetic components on human functioning with a lower-limb prosthesis. J Rehab R&D 2004;41(4):555–570.
40. Macfarlane PA, et al. Transfemoral amputee physiological requirements: comparisons between SACH foot walking and flex-foot walking. J Prosthe & Ortho 1997;9(4):138–143.
41. Nielsen DH, Shurr DG, Golden JC, Meier K. Comparison of Energy Cost and Gait Efficiency During Ambulation in Below-Knee Amputees Using Different Prosthetic Feet - a Preliminary Report. J Prosthet Orthotics 1989;1(1):24–31.
42. Heller BW, Datta D, Howitt J. A pilot study comparing the cognitive demand of walking for transfemoral amputees using the Intelligent Prosthesis with that using conventionally damped knees. Clin Rehab 2000;14(5):518–522.
43. Romo HD. Prosthetic knee. Phys Med Rehab Clin N Am 2000;11(3):595–607.
44. Taylor MB, Clark E, Offord EA, Baxter C. A comparison of energy expenditure by a high level trans-femoral amputee using the Intelligent Prosthesis and conventionally damped prosthetic limbs. Prosthet Ortho Int 1996;8:116–121.
45. Jia X, Zhang M, Lee WC. Load Transfer Mechanics Between Trans-Tibial Prosthetic Socket and Residual Limb - Dynamic Effects. J Biomech 2004;37(9):1371–1377.
46. Lee WC, Zhang M, Jia X, Cheung JT. Finite Element Modeling of the Contact Interface Between Trans-Tibial Residual Limb and Prosthetic Socket. Med Eng Phys 2004;26(8):655–662.
47. Winson CCL, Zhang M, Boone D, Contoyannis B. Finite-Element Analysis to Determine Effect of Monolimb Flexibility on Structural Strength and Interaction Between Residual Limb and Prosthetic. J Rehab R&D 2004;41(6a):775–786.
48. Phillips B, Zhao H. Predictors of Assistive Technology Abandonment. Assis Technol 5(1):1993; 178–184.
49. Scherer MJ. The change in emphasis from people to person: introduction to the special issue on assistive technology. Disabil Rehab 2002;24(1–3):1–4.
50. Marks LJ, Michael JW. Science, Medicine, and the Future: Artificial Limbs. BMJ 2001;323(7315):732–735.
51. Beil TL. Interface Pressures During Ambulation Using Suction and Vacuum-Assisted Prosthetic Sockets. J Rehab R&D 2002;39(6):693–700.
52. Michael JW. Modern Prosthetic Knee Mechanisms. Clin Orthop Relat Res 1999;361:39–47.
53. Buckley JG, Spence WD, Solomonidis SE. Energy Cost of Walking: Comparison of"Intelligent Prosthesis" With Conventional Mechanism. Arch Phys Med Rehabil 1997;78(3):330–333.
54. Twiste M, Rithalia SV, Kenney L. A Cam-Displacement Transducer Device for Measuring Small Two-Degree of Freedom Inter-Component Motion in a Prosthesis. Med Eng Phys 2004;26(4):335–340.
55. Lee WC, Zhang M, Jia X, Cheung JT. Finite Element Modeling of the Contact Interface Between Trans-Tibial Residual Limb and Prosthetic Socket. Med Eng Phys 2004;26(8):655–662.

See also BLIND AND VISUALLY IMPAIRED, ASSISTIVE TECHNOLOGY FOR; ENVIRONMENTAL CONTROL; LOCOMOTION MEASUREMENT, HUMAN; REHABILITATION AND MUSCLE TESTING.

MODELING OF PHYSIOLOGICAL SYSTEMS. See PHYSIOLOGICAL SYSTEMS MODELING.

MODELS, KINETIC. See TRACER KINETICS.

MONITORING IN ANESTHESIA

TARMO LIPPING
VILLE JÄNTTI
ARVI YLI-HANKALA
Tampere University of Technology
Pori, Finland

INTRODUCTION

Anesthesia is one of the most complex and mysterious phenomena in clinical work. The main feature of anesthesia is the loss of consciousness, which suggests its relatedness to sleep, epilepsy, and various kinds of brain trauma. In the case of anesthesia and sedation, consciousness is manipulated deliberately to prevent the patient from being aware of their state and the medical procedures carried through.

Recent decades have seen significant advancements in mapping various psychological functions to corresponding brain areas, however, the knowledge of the formation of human consciousness is still based on uncertain hypothesis. This complexity makes anesthesia monitoring extremely challenging.

This article first addresses the problem of anesthesia monitoring from the clinical, as well as from the physiological, point of view. The main emphasis is on monitoring anesthetic depth as this is the most discussed topic in anesthesia monitoring today. It starts with clinical indicators of anesthetic depth, gives an overview on the methods used in commercially available depth-of-anesthesia monitors, and describes some new algorithms proposed and evaluated for anesthesia electrocardiograms (EEG) monitoring in recently published works. Finally, the feasibility of monitoring brain function is argued using neurophysiological parameters like EEG, Auditory Evoked Potentials (AEPs), and so on, in the Intensive Care Unit (ICU) and the Emergency Room (ER).

ANESTHESIA AS A PROCESS AND A PROCEDURE

Anesthesia can be seen from the clinical point of view as a procedure, carried out according to certain protocol. From the physiological point of view anesthesia is a process evolving in the nervous system as the dose of an anesthetic agent increases.

Anesthesia as a Procedure

The goal of general anesthesia in the operating room (OR) is to render the patient unaware so that they do not feel pain during the surgery or recall the events afterward. It is also important that the patient does not react to surgical stimuli by movement. In the ICU, the goal of sedation is to keep the patient calm and painless. Too deep anesthesia causes prolonged awakening times after surgery in OR and longer treatment times in the ICU. The goals of anesthesia and sedation can be achieved by hypnotics (unconsciousness producing drugs), analgesics (antinociceptive drugs), and neuromuscular blocking agents. The choice of drugs is mainly made based on experience and clinical signs during the treatment.

Although general anesthesia is considered a safe procedure, various complications like postoperative nausea, vomiting, and pain are relatively frequent. The incidence of recall of events and awareness during anesthesia is rare ($\sim 0.1\%$), however, the consequences may be traumatic for the patient (1). Anesthesia-related mortality has decreased significantly during past decades having recently been estimated at 1 death per 200,000–300,000 cases of anesthesia (2–4).

Anesthetic agents can be divided into inhalation anesthetics (e.g., halothane and isoflurane) and intravenous anesthetics (e.g., thiopental and propofol). Intravenous drugs are becoming more popular as they are short acting, do not cause gas pollution, are easy to administer, and do not cause airway irritation. Desirable properties of anesthetic agents include rapid, smooth and safe induction of and emergence from anesthesia, no accumulation in the body, minimal effects on cardiovascular functions, no irritation to tissues and veins, low potential to hypersensitivity reactions.

Anesthesia as a Process

During the last decades, it has become obvious that different anesthetic and sedative agents produce their effects with different mechanisms, and therefore from the physiological point of view depth of anesthesia is a vague notion (5). It is more meaningful to talk about components forming the state we usually call anesthesia. These include amnesia, unconsciousness (hypnosis), antinociception, and neuromuscular blockade (paralysis). Different neurophysiological modalities should be used in order to assess these components. In the operating room, the patient is said to be anesthetized while the term sedation is used in the ICU. Some drugs, like propofol, are useful in producing both anesthesia and sedation at different concentrations, while some others are most useful as anesthetics or sedatives. There is evidence that anesthesia and sedation can be produced via different structures in the brainstem. Hypnosis and sedation cause similar changes in the EEG signal (described in the section changes in Neurophysiological Variables During Anesthesia). As all the available depth-of-anesthesia monitors are either fully or partly based on the EEG, the terms depth of hypnosis and depth of sedation are used depending on the corresponding clinical situation and are, in the context of available monitoring devices, roughly equivalent.

The action of anesthetics can be studied at various levels of neuronal function (6). The models underlying these studies can be divided into those operating at the molecular–cellular level and those explaining anesthetic processes at higher levels (generator models). State-of-the-art knowledge on the molecular and neuronal substrates for general anesthesia has recently been reviewed in Ref. 7. The model proposed by Flohr describes the action of anesthetics as disruption of computational processes dependent on the NMDA receptors (8). The activation state of these receptors in cortical neurons determines the complexity of representational structures that can be built-up in the brain and thus the level of consciousness. Steyn-Ross et al. performed numerical simulations of a single-macrocolumn model of the cerebral cortex and found that the effects of anesthetics can be modeled by cortical phase transition (9). Their simulations explain several trends in the EEG caused by anesthetic actions and predict the decrease in spectral entropy of the EEG signal with deepening anesthesia. This model has supported the development of the Entropy module, described in the section Recently Developed Methods, Applied in Commercial Anesthesia Monitors. Great cautiousness must be taken, however, in interpretation of the models operating at molecular level, because they include only a small part of the neurophysiological functions known to be involved in consciousness and generation of anesthesia-induced EEG patterns.

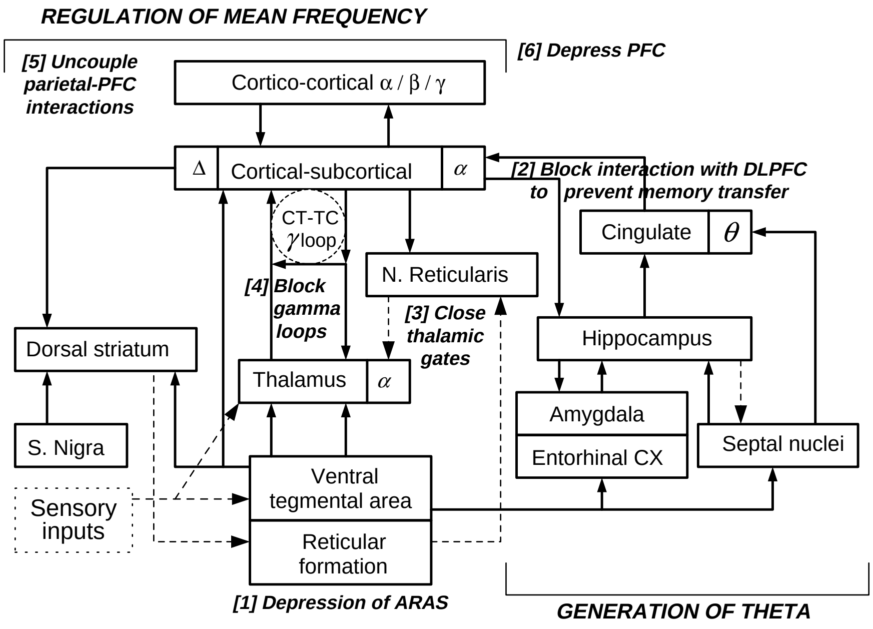


Figure 1. The Anesthetic Cascade model (redrawn with permission from: E. R. John and L. S. Prichep, "The anesthetic cascade: A theory on how anesthesia suppresses consciousness," Anesthesiology, Vol. 102, Fig. 11 (p. 468), 2005).

John et al. developed a higher level model based on a complex neuroanatomical system described in (10). This model, described and thoroughly discussed in volume 10 of Ref. 11, incorporates and explains an extensive bulk of results obtained from EEG, evoked potential and magnetic resonance imaging (MRI) image analysis, as well as laboratory studies. Loss of consciousness is described as the following cascade of events, called the Anesthetic Cascade by the authors, involving various brain regions (Fig. 1) (6): (*1*) depression of the brainstem; (*2*) depression of mesolimbic-dorsolateral prefrontal cortex interactions leading to blockade of memory storage; (*3*) inhibition of the nucleus reticularis of the thalamus, resulting in closure of thalamic gates (seen as increasing θ rhythm in the EEG); (*4*) blocking of thalamocortical reverberations (γ loop) and perception; (*5*) uncoupling of parietal-frontal transactions (coherence in γ frequency band decreases); (*6*) depression of prefrontal cortex to reduce awareness (increase in frontal θ and δ rhythm).

Definitions of the EEG rhythms used in the description of the Anesthetic Cascade are given in Table 1.

The model by John et al. underlies the Patient State Index for depth-of-anesthesia monitoring, described in the section Recently Developed Methods, Applied in Commercial Anesthesia Monitory. Although the Anesthetic Cascade model covers a large set of neurophysiological functions, it does not explain patterns like burst suppression, for example.

Table 1. Definition of the EEG Rhythms[a]

EEG Rhythm	Frequency Range, Hz
Delta (δ)	< 4
Theta (θ)	4–8
Alpha (α)	8–12
Beta (β)	12–25
Gamma (γ)	25–50

[a] Exact frequencies may vary slightly from source to source.

MONITORING ADEQUACY OF ANESTHESIA

Clinical Indicators and Measures of Anesthetic Depth

The verb monitor originally means to check systematically or to keep watch. Thus, monitoring actually does not necessarily involve medical equipment, but refers also to clinical inspection. As clinical indicators of anesthetic depth are often used as a reference for automated depth-of-anesthesia monitoring methods, they are shortly described here.

In the case of inhalation anesthetics, drug concentration can be monitored by measuring the partial pressure of the anesthetic in exhaled air (end-tidal concentration). Due to the variation of the potency of anesthetic agents, a universal unit of minimum alveolar concentration (MAC) has been applied. The value 1 MAC is the partial pressure of an inhaled anesthetic at which 50% of the unparalyzed subjects cease to express protective movement reaction to skin incision. The primary rationale behind the development of the term MAC was the need to compare the potency of different volatile anesthetics, not the effort to monitor the anesthetic state of an individual patient.

For intravenous anesthetics, no such direct measure can be derived and the effect of anesthetics can be estimated using pharmacokinetic models (effect-site concentration). In this case, the accuracy of the estimate depends on the adequacy of the model. If all subjects would react to anesthetics in exactly identical ways these concentration measures would provide a perfect indicator of adequacy of anesthesia. However, there is an intersubject variability in the effect of anesthetics, and therefore other indicators are needed.

Clinical indicators of the adequacy of surgical anesthesia can be divided into those measuring hypnosis and those measuring nociceptive–antinociceptive balance. The indicators measuring hypnosis include pupillary light reflex, tested by allocating a flashlight to one eye and observing both pupils for constriction; corneal reflex,

Table 2. Ramsay Score for Assessment of Level of Sedation[a]

Score	Clinical Status
1	Patient anxious and/or agitated
2	Patient cooperative
3	Patient responds to commands only
4	Brisk response
5	Sluggish response
6	No response to loud auditory stimulus

[a]See Ref. 13.

tested by applying a wisp of cotton wool to the cornea or by electrical stimulation using special electrodes (12); Eyelash reflex, tested by brushing the eyelashes with a moving object or by electrical stimulation; loss of counting, tested by letting the subject count slowly as long as they can from the onset of infusion–injection; syringe dropping, tested by letting the subject hold a syringe between their thumb and forefinger as long as they can; loss of obeying verbal commands.

The indicators measuring nociceptive–antinociceptive balance include avoidance reaction to nociception. This is mainly a spinal reflex, however, it correlates well with the concentration of most anesthetics; electrical tetanic stimulation, applied using needle electrodes or adhesive skin electrodes to the upper or lower limb; autonomic nervous system—mediated reactions or motor reactions to laryngoscopy and endotracheal intubation. This is a natural stimulus in many clinical situations in the operating room.

These indicators test the functioning of different neural pathways and their applicability depends on the anesthetic used. For example, ketamine leaves corneal and pupillary light reflexes intact.

For more graded and standardized clinical assessment of sedation and hypnosis, several scoring systems have been developed. Probably the most widely used such systems are the Ramsay score (Table 2) and the Observer's Assessment of Alertness and Sedation Scale (OAAS; Table 3). These scoring systems are developed for use in the ICU as they include scores for agitated states and cover mainly lighter levels of anesthesia. Therefore, they do not necessarily indicate the adequacy of anesthesia for surgical procedures. Also, the assessment obtained using these scoring systems is subjective.

Table 3. OAAS Score for Assessment of Level of Sedation[a]

Score	Clinical Status
5	Responds readily to command spoken in normal tone
4	Lethargic response to command spoken in normal tone
3	Lethargic response to command spoken loudly and repeatedly
2	Appropriate response to loud tone and mildly painful stimulus
1	Appropriate response to loud tone and moderately painful stimulus
0	No response

[a]See Ref. 14.

Changes in Neurophysiological Variables with Deepening Anesthesia

All the commercial monitors of hypnosis employ the EEG signal. Although different anesthetic agents induce specific features and patterns in the EEG, certain common trends in signal properties with deepening anesthesia can be seen. At subanesthetic levels, several agents produce oscillations at beta frequency range, sometimes called beta buzz. This activity is seen dominantly in the frontal brain areas. With increasing anesthetic concentrations, the activity becomes more widespread, decreases in frequency and increases in amplitude. Around concentrations, causing the subjects to stop responding to stimuli (1 MAC for inhalation anesthetics), the EEG activity slows further and high amplitude theta and delta waves occur. With further increasing concentration, the burst-suppression (BS) pattern occurs, finally turning into continuous suppression. The dynamics of this pattern, as well as the waveforms of bursts, varies for different anesthetic agents (Fig. 2). Several anesthetic agents tend to induce epileptiform seizure activity in patients with a prior history of seizures and even in subjects with no previous history of seizures (15,16).

In addition to the EEG signal, AEPs have been used for anesthesia monitoring. The latency of early cortical responses Pa and Nb increases and the amplitude decreases with deepening anesthesia (17). Also, late cortical responses to auditory stimuli, specifically the amplitude and latency of the N100 peak have been found to improve the assessment of the level of consciousness in ICU patients (18). A commercially available brain monitor, the AEP Monitor/2 by Danmeter A/S, combines AEPs with EEG parameters to calculate the cAAI index (see the next section).

In most commercially available monitoring devices, the EEG signal is obtained from the electrodes placed at the forehead. This makes the recording procedure easier in clinical situations. The electrodes tend to pick up frontal EMG, which is an artifact from the point of view of the EEG signal but may be used as a valuable indicator of nociception in light anesthesia (19). The EMG component of the measurement is either explicitly or implicitly incorporated into most of the available monitoring devices (see the section Discussion).

Another neurophysiological variable proposed for anesthesia monitoring is the respiratory sinus arrhythmia (RSA) component of the heart rate (HR) signal (20). Although potentially valuable addition to the assessment of the level of consciousness, this variable has not made its way to anesthesia monitoring devices to date.

Short History of Brain Monitoring in Anesthesia

Since the first measurements of human electroencephalogram, performed by Hans Berger in 1920s, this modality has been applied to studying the effects of various drugs, including anesthetics. The emergence of microprocessors and digital techniques for signal analysis opened new perspectives for anesthesia monitoring.

The first commercial brain monitoring device based on digital signal analysis was the *Cerebral Function Analyzing Monitor* (CFAM1), developed in 1975 by Prior and

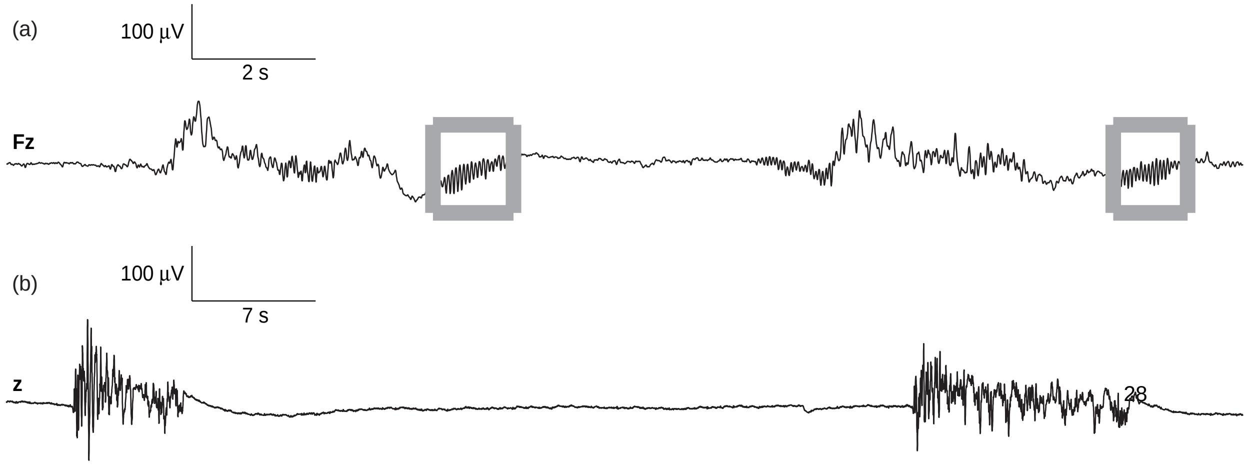


Figure 2. Samples of BS pattern in EEG during deep propofol (a) and sevoflurane (b) anesthesia. Detection of BS suppression and calculation of BS ratio is an important part of all modern depth-of-hypnosis monitors. The pattern varies significantly among anesthetic drugs. In the case of propofol anesthesia, spindles can be observed (marked by boxes in the figure). Note that the scale of the time axes is different for upper and lower curves.

Maynard (21). This device used the Motorola 6808 8-bit microprocessor. The display of the CFAM1 was divided into two sections, one showing the 10th and 90th percentile as well as the mean of the EEG amplitude distribution while the other showing the percentage of weighted (prewhitened) EEG activity per herz in the beta, alpha, theta, and delta frequency bands (Fig. 3). In addition, muscle activity, EEG suppression ratio, and electrode impedance were displayed. An important feature of the CFAM1 was the possibility of monitoring averaged evoked potentials. Since the introduction of CFAM1, the CFAM family of brain monitors has been continuously developed with the recently introduced CFAM4 being the latest member of this product family. Comprehensive list of publications referring to the CFAM family can be found at www.cfams.-com/references/a4a.htm.

In 1982 Datex-Ohmeda (Helsinki, Finland) introduced its first EEG monitor for anesthesia, *the Anesthesia Brain Monitor* (ABM). Like in most of the later monitoring devices, the location of the EEG electrodes in the ABM monitor was on the forehead. The monitor displayed the root-mean squared (rms) value of the EMG and the RMS, as well as the zero-crossing frequency of the EEG signal. The EMG and EEG signals were obtained from the same electrodes–bandpass filter of 65–300 Hz was applied to obtain the EMG while frequencies 1,5–25 Hz were used to obtain the EEG. The ABM monitor is described in (22).

At the beginning of 1990s Thomsen et al. took a different approach to anesthesia monitoring in their *Advanced Depth of Anesthesia Monitor* (ADAM) (23). They divided the signal into consecutive 2 s segments, applied a prewhitening filter, and derived 11 parameters: the rms value and 10 correlation coefficients from each segment. Either the values of the first 10 autocorrelation lags or the coefficients of the 10th-order autoregressive model were suggested as features. To create a set of reference classes, an unsupervised repetitive hierarchical cluster analysis was applied to the data bank of preannotated recordings of halothane and isoflurane anesthesia. Six clusters were defined, corresponding to anesthetic levels from drowsiness to very deep anesthesia. The classification was adjusted according to the anesthetic agent used. Burst-suppression was detected separately and the suppression ratio in 2 s segments was incorporated into classification. Anesthetic depth was displayed as the class probability histogram: A plot where each line represented the clusters

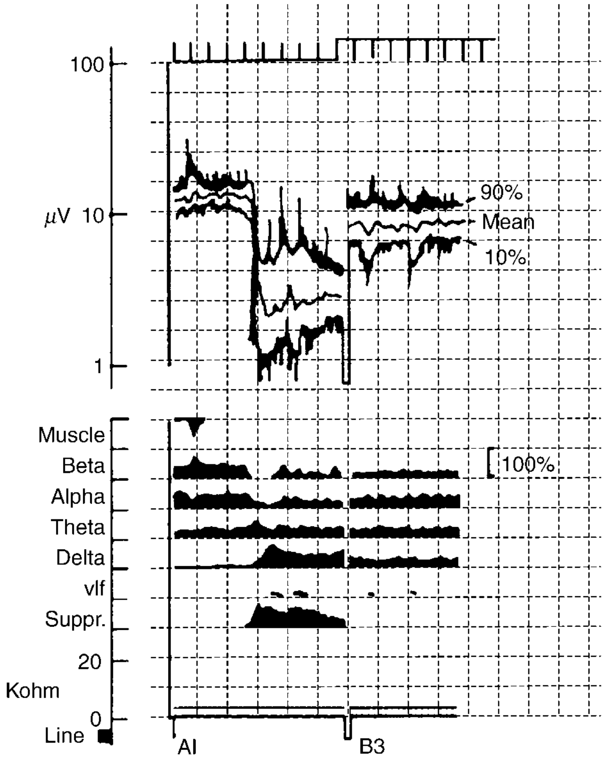


Figure 3. Layout of the screen of the CFAM1 monitor (with permission from D. Maynard).

obtained for 10 s period of the recording. The clusters were color coded. In spite of its advanced approach, ADAM was never implemented in a commercial anesthesia monitoring device.

Recently Developed Methods, Applied in Commercial Anesthesia Monitors

The *Bispectral Index Score* (BIS), developed by Aspect Medical Systems Inc. in 1997, marked a breakthrough in anesthesia monitoring. The output of the BIS monitor is a single number between 0 and 100 achieved by combining in a nonlinear fashion from the following parameters (24): relative beta ratio calculated in spectral domain as $\log\left(\frac{P_{30-47}}{P_{11-20}}\right)$, where P_{30-47} and P_{11-20} denote signal power in frequency ranges 30–47 and 11–20 Hz, respectively; SynchFastSlow measure calculated in bispectral domain as $\log\left(\frac{B_{0.5-47.0}}{P_{40.0-47.0}}\right)$, where $B_{0.5-47.0}$ and $B_{40.0-47.0}$ denote the sum of magnitudes of the bispectrum values in the corresponding frequency ranges; BS ratio. Bispectrum (the third-order spectrum), is defined as the two-dimensional (2D) Fourier transform (FT) of the third-order cumulant sequence $c_3(k_1, k_2)$ of the signal:

$$B(\omega_1, \omega_2) \overset{\mathrm{FT}}{\leftrightarrow} c_3(k_1, k_2) \tag{1}$$

If the direct current (dc) component of the signal has been removed (as is usually the case), $c_3(k_1, k_2)$ equals to the third order moment sequence $m_3(k_1, k_2)$, defined as:

$$m_3(k_1, k_2) = \varepsilon\{s(n)s(n+k_1)s(n+k_2)\} \tag{2}$$

where $\varepsilon\{\cdot\}$ denotes expected value. Overview on the estimation of higher order spectra can be found in (25).

The weighting of the three parameters forming the BIS depends on signal properties and is not disclosed. In light anesthesia, relative beta ratio is dominating while SynchFastSlow measure becomes more important with deepening anesthesia. The function combining the parameters was developed empirically, based on thousands of EEG records. An important part of BIS is its careful artifact rejection scheme, dealing with heartbeat artifacts, eyeblinks, wandering baseline, muscle artifacts, and so on BIS has become very popular among anesthesiologists; the bulk of literature dealing with the behavior of BIS in various clinical situations, discussing its advantages as well as disadvantages, incorporates more than 1000 papers. Comprehensive bibliography can be found on the web-pages of Aspect Medical Systems Inc.

At the beginning of this decade Physiometrix Inc. brought to market the PSA 4000 depth-of-hypnosis monitor, based on the *Patient State Index* (PSI) (26). The development of the PSI was based on a library of 20,000 cases of EEG records. In addition, a library of surgical cases, a library of artifacts and results from volunteer studies (for calibration), were used. In PSI, the EEG signal is measured from four electrodes: Fp1, Fpz, Cz, and Pz, with the reference at linked ear electrodes. Signal analysis is based on power in standard EEG frequency bands (see Table 1) and incorporates the calculation of the following parameters: absolute power gradient between Fp1 and Cz leads in the gamma frequency band (25–50 Hz); absolute power changes between Fpz and Cz leads in beta (12–25 Hz) and between Fpz and Pz; leads in alpha (8–12 Hz) frequency bands; total spectral power (0.5–50 Hz) at the Fp1 lead; mean frequency of the total spectrum at Fpz lead; absolute power in delta frequency band (0.5–4 Hz) at Cz; relative power at Pz lead in slow delta frequency band;

The calculated parameters go through a mathematical transformation that guarantees their Gaussian distribution in order to be rescaled into the Z-score (Fig. 4). The Z-score sets the calculated parameters into relation with the parameter values obtained for reference population giving the percentage of the reference population that lies more standard deviations away from the mean than the calculated parameter (6). The Z-scored parameters are fed into discriminant analysis with adaptive discriminant functions. EEG suppression is detected separately: The suppression ratio is included in the discriminant analysis. The discriminant analysis yields the PSI index: a scalar between 0 and 100 with higher level of hypnosis corresponding to lower PSI value.

The *Narcotrend* anesthesia monitoring system was developed by a German group and first introduced in 2000 (27,28). This system has its roots in sleep analysis: A five-stage sleep scoring system was further developed into a system of 6 stages and 14 substages for level-of-hypnosis monitoring. These stages are mapped to a scale of 0–100 in the Narcotrend algorithm. The EEG signal is derived from one or two electrodes; the most common electrode location is on the forehead, however, according to the authors other electrode locations are possible. The signal is sampled at 128 Hz and prefiltered using lower and upper cutoff frequencies of 0.5 and 45 Hz, respectively. The principal idea underlying the method is similar to that of the PSI: Several variables calculated from the EEG signal are fed to discriminant analysis with separate detection of BS (Fig. 5). The variables are classified as time- and frequency-domain ones and contain signal power, autoregressive coefficients, relative power in standard EEG frequency bands, median frequency (the frequency dividing the signal spectrum into two parts of equal energy), spectral edge frequency (SEF95, the frequency below which 95% of signal energy is contained), and spectral entropy. The algorithm also contains plausibility check to ensure that the signal segment is actually similar to a typical EEG sample of corresponding stage and to detect patterns in the EEG signal untypical for general anesthesia (e.g., epileptic activity). The detailed algorithm of the Narcotrend index is proprietary.

Another EEG-based depth-of-anesthesia monitoring device is the recently introduced *M-Entropy* module for the Datex-Ohmeda S/5 anesthesia monitor. As the name indicates, the method is based on the idea that the entropy of the EEG signal decreases with deepening anesthesia. Signal entropy can be defined and calculated in many different ways (see also the next section) of which spectral entropy is employed in the M-Entropy module. Spectral entropy in the frequency range f_1–f_2 is expressed as

$$S(f_1, f_2) = \sum_{f_i=f_1}^{f_2} P_n(f_i) \log \frac{1}{P_n(f_i)} \tag{3}$$

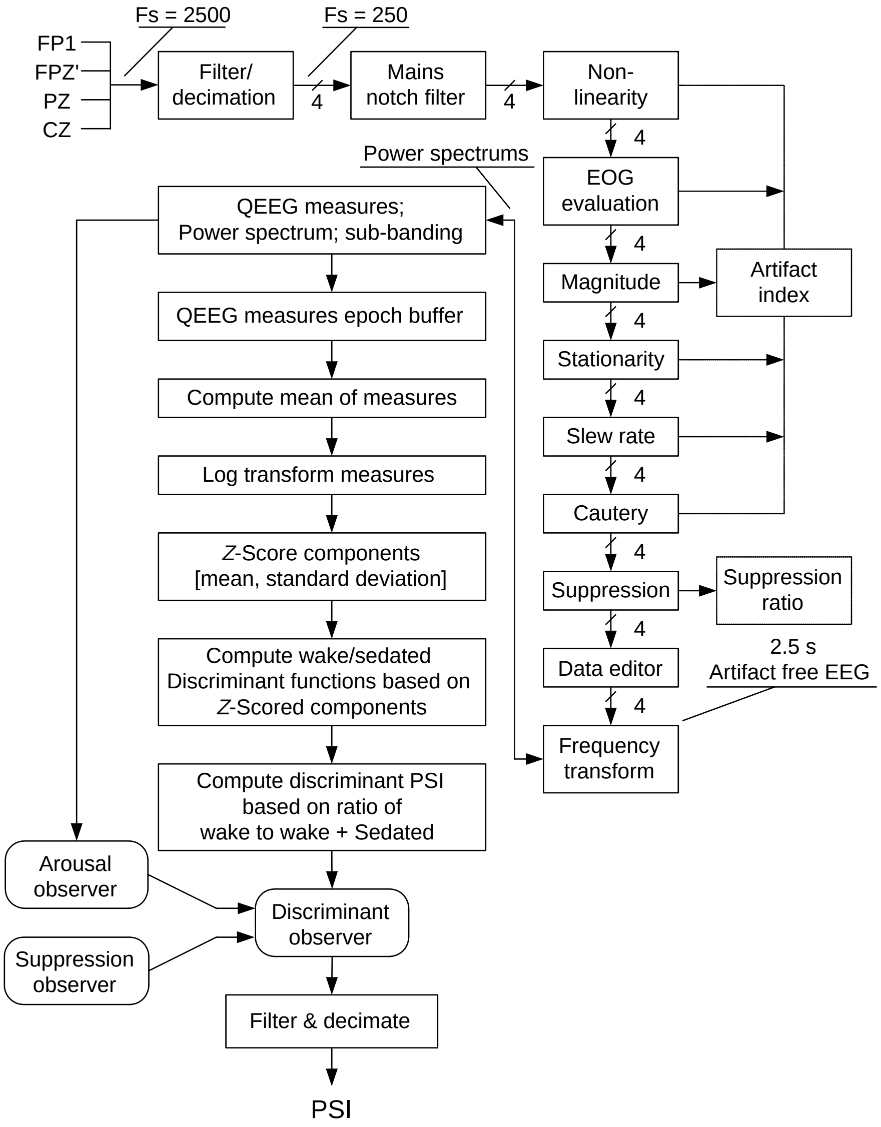


Figure 4. Schematic of the calculation of the PSI index. (Redrawn with permission from D. R. Drover et al., “Patient State Index: Titration of delivery and recovery from propofol, alfentanil, and nitrous oxide anesthesia,” Anesthesiology, Vol. 97, Fig. 3 (p. 88), 2002).

where $P_n(f_i)$ is the normalized power spectrum of the signal. $S(f_1, f_2)$ is again normalized by log $N(f_1, f_2)$, where $N(f_1, f_2)$ is the number of frequency components in the range f_1–f_2, to give a value between 0 and 1. In the original version of the device, the analysis was performed on a single EEG channel measured from the forehead. In this derivation, muscle activity dominates over the EEG at frequencies higher than ~30 Hz. The algorithm of the M-Entropy module, like that of the early ABM-monitor by Datex-Ohmeda, employs these high frequency components to detect the early response of the patient to nociceptive stimuli. This is done by calculating spectral entropy over two frequency ranges: 0.8–32 Hz (called state entropy) and 0.8–47 Hz (called response entropy). The difference between these two entropies indicates the contribution of the EMG component to the response entropy. As in the other described monitors, BS is detected separately. The details of the algorithm (variable window length, obtaining the output value in the case of BS, etc.) are described in (29).

The Danmeter AEP Monitor/2 (further development of the A-Line monitor) employs the composite AAI Index, combining the middle latency auditory evoked potentials in 20–80 ms latency range, calculated from the 25–65 Hz bandpass filtered signal, with spontaneous EEG. The purpose of combining the two modalities is to get a better response to the lightening of hypnosis due to, for example surgical stimuli (achieved by usind AEPs) while retaining sensitivity during deep anesthesia (achieved by using the EEG). The schematic of the cAAI index calculation is presented in Fig. 6. Using evoked potentials poses a problem in on-line monitoring due to the long delay needed for obtaining the averaged response. This problem has been solved in the cAAI calculation by applying the autoregressive model with exogenous input (the ARX model). The ARX model enables to calculate the response to stimuli based on the average of 18 sweeps using the average of 256 sweeps as a reference. The algorithm is described in detail in (30) and compared with conventional evoked potential averaging techniques in Ref. 31. In addition to AEPs, the cAAI index incorporates logarithmic EEG power ratio $[\log(P_{30-47}/P_{10-20})]$ and the burst suppression ratio. The EMG is extracted and monitored

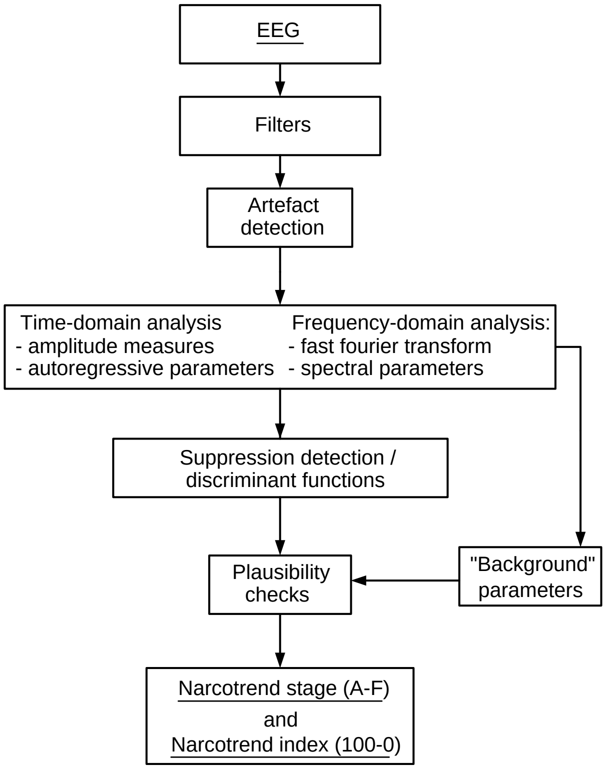


Figure 5. Schematic of the calculation of the Narcotrend index (with permission from B. Schultz).

separately based on the 65–85 Hz bandpass filtered signal.

A somewhat different concept of anesthesia monitoring has been used in the *Cerebral State Monitor* (CSM; Danmeter A/S, Odense, Denmark) and the *SNAP* monitor (Everest Biomedical Instruments Inc). These monitoring devices come in the form of a handheld wireless PDA-type tool, convenient to use in a clinical situation. The CSM displays the Cerebral State Index, calculated based on the 6–42 Hz bandpass filtered EEG, the EMG component calculated from the same signal, but in 75–85 Hz frequency range, as well as the burst suppression ratio. The algorithm of the second version of the SNAP index is described in (32). Two variables, the low frequency variable *LF* (0.1–40 Hz) and the high frequency variable *HF* (80–420 Hz) are derived from a single frontal EEG channel. The HF and LF are scaled to fit into the intervals 0.0–1.0 and 0.0–100, respectively. The SNAP index is expressed as SI = 100—(HF*LF); thus the index can be thought of as the reversed version of HF-modulated LF.

New Parameters Proposed for Monitoring Anesthetic Depth

In spite of the large selection of available methods, new parameters for quantifying depth of hypnosis are being proposed continuously. This is mostly due to the following reasons: the variety of procedures and combinations of drugs in surgical anesthesia is wide. No method performs well in all cases; monitoring in anesthesia is closely related to monitoring brain dysfunction and detection of brain ischemia and hypoxia: important tasks faced in cerebral function monitoring in the ICU and emergency room (see the section Monitoring Outside the Operating Theater). The available depth-of-hypnosis monitors are generally not suitable for these applications; the neurophysiological basis of consciousness is still an unsolved problem: applying modern signal analysis tools to neurophysiological measurements during anesthesia can hopefully offer new insight to the problem.

Several groups have recently published studies on the behavior of various complexity–entropy measures during

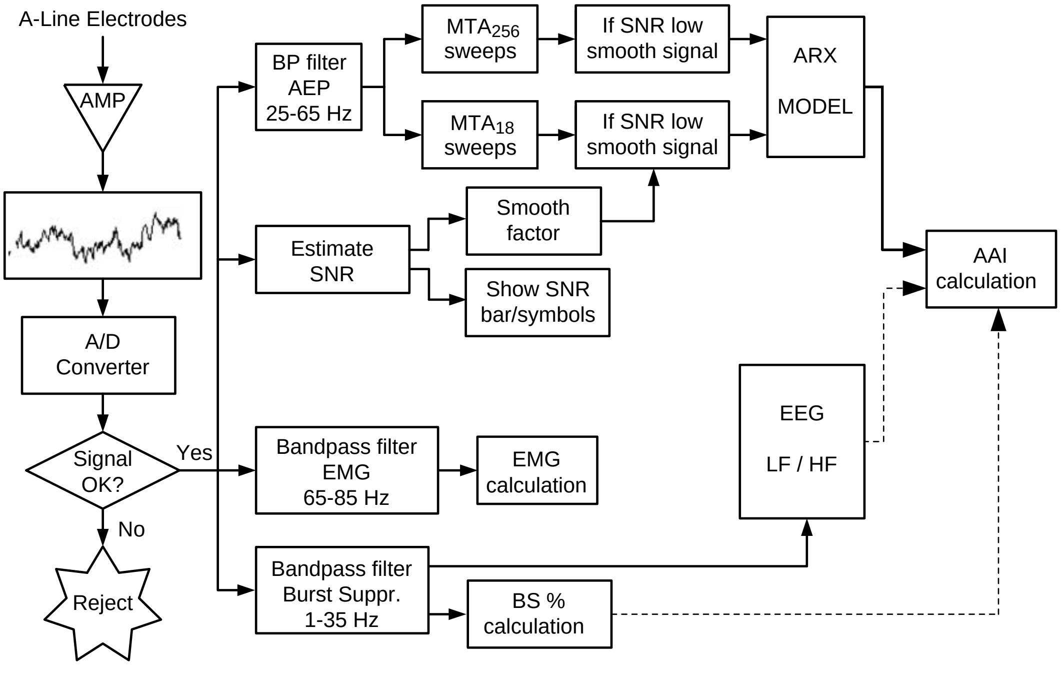


Figure 6. Schematic of the calculation of cAAI index (with permission from E. W. Jensen).

anesthesia and sedation. These measures come from different signal analysis frameworks.

Correlation dimension is a measure for quantifying the behavior of chaotic signals in the phase space (33). The signal s of finite length N is divided into $N{-}m{+}1$ time series: $s_m(i) = \{\mathrm{s}(i), \mathrm{s}(i+1), \ldots, \mathrm{s}(i+m{-}1)\}$, where m is the embedding dimension. After that, for each i the quantity $C_{\mathrm{i}}^{\mathrm{m}}(r)$ is calculated:

$$C_i^m(r) = \frac{\text{number of such}\, j\, \text{that}\, d[S_m(i), S_m(j)] \le r}{N - m + 1} \quad (4)$$

where the distance d between the phase space vectors $s_m(i)$ and $s_m(j)$ is defined as

$$d[S_m(i), S_m(j)] = \max_{k=1,2,\ldots,m} (|s(i+k-1) - s(j+k-1)|) \quad (5)$$

Correlation dimension D can be estimated as

$$D = \frac{d\log(C^m(r))}{d\log(r)} \quad (6)$$

where $C^m(r) = \sum_i C_i^m(r)/N - m + 1$. Although EEG cannot be considered strictly chaotic, except in the case of some abnormal conditions, this measure has, for example, been found to have good correlation with the end-site concentration of sevoflurane (34).

Probably the most intensively studied complexity/entropy measure for the assessment of depth of hypnosis is Approximate entropy (ApEn). In general, entropy measures information-richness, regularity and randomness of a signal. The intuitive idea behind anesthesia monitoring using signal entropy is that with deepening anesthesia EEG becomes more regular and its entropy decreases. Approximate entropy, like correlation dimension, is calculated in the phase space. First, $\phi^m(r)$ is defined based on $C_i^m(r)$ in Eq. 4 as

$$\Phi^m(r) = \frac{1}{N - m + 1} \sum_{i=1}^{N-m+1} \log C_i^m(r) \quad (7)$$

Approximate entropy is then defined as

$$ApEn(m,r) = \Phi^m(r) - \Phi^{m+1}(r) \quad (8)$$

Approximate entropy has been studied and compared to other methods as an indicator of anesthetic depth in (35–37).

The classical entropy measure, introduced for information theory by Claude Shannon in 1948 (38), the Shannon entropy (ShEn), is calculated as $\mathrm{ShEn} = -\sum_i p_i \log p_i$, where p_i is the probability that signal amplitude obtains the range of values a_i. In practice, ShEn can be estimated based on the histogram of the values of signal samples, and therefore long signal segments are needed to achieve smooth histograms. An important property of Shannon entropy is that signal samples are considered as independent trials of some experiment, taking no notice on the time order of the samples. Signals having equal probability for all possible amplitude values have the highest Shannon entropy. In Ref. 39, it has been found that Shannon entropy of the EEG recorded between frontopolar electrodes increases with increasing concentration of desflurane: A behavior opposite to other entropy measures.

Other measures of the EEG, found to correlate well with depth of hypnosis, include Lempel–Ziv complexity and Higuchi fractal dimension (35,40). Lempel–Ziv complexity is calculated by transforming the signal into symbols and calculating the reoccurrence rate of these symbols (41). Higuchi fractal dimension is calculated as the average rate of increase in the difference of signal amplitude values as the separation between the samples increases in logarithmic scales (42).

These studies demonstrate that although different measures of entropy or complexity quantify different phenomena, many of them may correlate with concentrations of selected anesthetics when electrode positions and signal bandwidth are selected properly.

MONITORING OUTSIDE THE OPERATING THEATER

Development of digital EEG equipment, increase in processing speed and memory capacity, and advancements in telecommunication technology have made cerebral function monitoring feasible in ICU and ER. Brain monitoring in ICU and ER has much in common with monitoring in anesthesia as the changes in the EEG caused by intoxication, metabolic disturbances and brain ischaemia are similar to those induced by general Anesthesia. Also, in the ICU the assessment of depth of sedation is desirable. The advantages offered by EEG monitoring in the ICU are based on the following findings (43): EEG is tightly linked to cerebral metabolism; EEG is sensitive to brain ischemia and hypoxia; EEG detects neuronal dysfunction at a reversible stage; EEG detects neuronal recovery when clinical examination cannot; continuous EEG provides dynamic information; EEG provides useful information about cerebral topography.

However, from the monitoring point of view, the situation in ICU and ER is a lot more complicated compared to that of OR. The patients may need various medication having effect on the EEG signal and misleading automated EEG analysis, the clinical situation of the patients is often complex, and the surrounding is hostile for interference-sensitive equipment. In ICU, recordings often need to last for several days and nights without disturbing the normal care of the patient. In ER, the EEG recording equipment needs to be extremely flexible and easy-to-use. In both situations the interpretation of the recordings poses a problem as no experienced EEG readers are usually around. The solution to the last problem is the usage of telecommunication protocols to transfer the data for interpretation.

Although the above described depth-of-anesthesia monitoring methods are sometimes applied to sedation monitoring and even to the detection of brain dysfunction in ICU, their performance in this situation is questionable. It is difficult to differentiate between the effects of hypoxia, ischemia and sedative drugs. The importance of having the underlying raw EEG signal available for review to confirm the significance of any trends and changes suggested by automatic analysis methods, especially in complex situations like ICU, has been repeatedly emphasized (44,45). A comprehensive brain monitor for ICU, especially for neuroscience ICU, should also be able to detect epileptic patterns in EEG and desirably have the option for synchronous video recording (46). All this suggests that an adequate brain monitor for ICU or ER should be a much more complex device than today's depth-of-anesthesia monitors.

DISCUSSION

Several considerations are appropriate concerning the available commercial depth-of-anesthesia monitors. First, different modalities should be used to assess the different components of anesthesia (see the section Anesthesia as a Process). Selecting EEG and AEPs as the basis for the assessment, the primary component of anesthesia considered would be hypnosis.

But even in this case there still remain other physiologically separate end-points like subcortically controlled reactions to nociceptive input (e.g., autonomic reactions), increased muscle tone, and movement response to surgery. Adding the fact that there are many anesthetic agents of different cell-level actions and that the interplay of hypnotic and antinociceptive medication modulates the anesthetic state (47), we are left with a complex situation that makes the comparison of the available algorithms for anesthesia monitoring a real challenge.

Another difficulty in comparing the results obtained with different methods is posed by the frequency band used for the calculation. All the commercial methods operate at least partly in the frequency domain although BIS applies third-order spectrum and in the Entropy module a nonlinear transformation follows the calculation of the power spectrum. For anesthesia, monitoring frequency domain can roughly be divided into the following physiologically meaningful areas: δ (and partly θ) frequencies, indicative of pre-BS deep anesthesia (~ 0.5–6 Hz); α and β frequencies; the EMG component, overlapping with the EEG and extending to > 100 Hz.

The devices differ in the usage of δ frequencies and in the way the EMG component is incorporated. While most of the methods employ frequency band starting from 0.1 Hz (SNAP)–0.8 Hz (Entropy), the Cerebral State Index and cAAI by Danmeter do not make use of δ rhythms. Several devices like the A-2000 monitor by Aspect Medical Systems Inc., the AEP Monitor/2 and the Cerebral State Monitor by Danmeter as well as the PSA 4000 monitor by Physiometrix Inc. display the EMG power separately from their corresponding depth-of-anesthesia indices. The frequency band the EMG component is obtained from varies from device to device, falling into the range from 65 to 110 Hz. The SNAP index, the Entropy module and the Narcotrend index incorporate the information on EMG activity into their depth-of anesthesia indexes in different ways. In SNAP, the high frequency band used is 80–420 Hz, while the other two monitors use frequencies up to 47 Hz. The various entropy–complexity measures proposed for the assessment of anesthetic depth are sensitive to the prefilter settings as well (40). Thus it can be concluded that while comparing the performance of various algorithms, the following matters should be considered: the properties of the algorithm itself, the frequency band of the EEG signal it employs, and the location of the EEG electrodes.

In the future, it seems to be inevitable that brain monitoring becomes more common in ICU and emergency room. There is a compromise between the simplicity of the presentation of the output and versatility of the method. Monitoring such complex phenomenon as anesthesia by a single number is clearly an oversimplification. On the other hand, a device requiring sophisticated configuration and displaying a lot of parameters difficult to interpret gets easily rejected by clinicians. Connecting the algorithms to physiological models would certainly help the interpretation of the monitor output. Future will show if any of the new approaches such as measures of signal complexity find their way into the commercial devices. Operating in the frequency domain has the advantage of long-term experience in EEG analysis by means of frequency domain methods. Another advantage is the solid signal processing theory of frequency analysis. On the other hand, the theory of nonlinear systems is developing rapidly having made itsway to physiological signal analysis in various applications.

BIBLIOGRAPHY

1. Myles PS, et al. Patient satisfaction after anesthesia and surgery: Results of a prospective survey of 10811 patients. Br J Anaesth 2002;84:6–10.
2. Committee on Quality of Health Care in America IoM. In: Kohn L, Corrigan J, Donaldson M, editors. To Err Is Human: Building a Safer Health System. Washington: National Academy Press 1999;241.
3. Chopra V, Bovill J, Spierdijk J. Accidents, near accidents and complications. Br J Anaesth 1978;50(10):1041Ü6 36.
4. Lagasse RS. Anesthesia safety: Model or myth? A review of the published literature and analysis of current original data. Anesthesiology 2002;97:1609–1617.
5. Kissin I. General anesthetic action: An obsolete notion? Anesthol Analg 1993;76:215–218.
6. John ER, Prichep LS. The anesthetic cascade: A theory on how anesthesia suppresses consciousness. Anesthesiology 2005; 102:447–471.
7. Rudolph U, Antkowiak B. Molecular and neuronal substrates for general anaesthetics. Nature Rev Neurosci 2004;5:709–720.
8. Flohr H, Glade U, Motzko D. The role of the NMDA synapse in general anesthesia. Toxicol Lett 1998;100-101:23–29.
9. Steyn-Ross DA, Steyn-Ross ML, Wilcocks LC, Sleigh JW. Toward a theory of the general-anesthetic-induced 24 phase transition of the cerebral cortex. II. Numerical simulations, spectral entropy, and correlation times. Phys Rev E 2001; 64:011918 (12 p).
10. Hughes JR, John ER. Conventional and quantitative electroencephalography in psychiatry. J Neuropsychiat Clin Neurosci 1999;11:190–208.
11. Volume 10: Conciousness and Cognition 2001.
12. Mourisse J, et al. Electromyographic assessment of blink and corneal reflexes during midazolam administration: Useful methods for assessing depth of anesthesia? Acta Anaesthesiol Scand 2003;47:593–600.
13. Ramsay M, Savege T, Simpson B. Controlled sedation with alphaxolene/alphadalone. Br J Med 1974;2:656–659.
14. Chernik D, et al. Validity and reliability of the observer's assessment of alertness/sedation scale: Study with intravenous midazolam. J Clin Psychopharmacol 1990;10:244–251.

15. Yli-Hankala A, et al. Epileptiform electroencephaogram during mask induction of anesthesia with sevoflurane. Anesthesiology 1999;91:1596–1603.
16. Vakkuri A. Effects of Sevoflurane Anesthesia on EEG Patterns and Hemodynamics. Ph.D. dessertation, University of Helsinki, Finland, 2000.
17. Thornton C, et al. The auditory evoked response as an indicator of awareness. Br J Anaesthology 1989;63:113–115.
18. Yppärilä H. Depth of Sedation in Intensive Care Patients: A Neurophysiological Study. Ph. D. Dessertation, University of Kuopio, Finland, 2004.
19. Paloheimo M. Quantitative surface electromyography (qEMG): Applications in anaesthesiology and critical care, Ph. D. dissertation Acta Anaesthesiology; Scandinavian. Copenhagen. Munksgaard; 1990; Vol. 34, (Suppl. 93),.
20. Wang DY, Pomfrett CJD, Healy TEJ. Respiratory sinus arrhythmia: A new, objective sedation score. Br J Anaesthology 1993;71:354–358.
21. Maynard D. Development of the CFM: The cerebral function analyzing monitor (CFAM). Ann Anaesthology Francaise 1979;20:253–255.
22. Edmonds HL, Paloheimo M. Computerized monitoring of the EMG and EEG during anesthesia. An evaluation of the anesthesia and brain activity monitor (ABM). Int J Clin Monit Comp 1985;1:201–210.
23. Thomsen CE, Rosenfalck A, Norregaard-Christensen K. Assessment of anaesthetic depth by clustering analysis and autoregressive modeling of electroencephalograms. Comput Methods Progr Biomed 1991;34:125–138.
24. Rampil IJ. A primer for EEG signal processing in anesthesia. Anesthesiology 1998;89:980–1002.
25. Nikias CL, Petropulu AP. Higher Order Spectra Analysis. Englewood Cliffs (NJ): PTR Prentice Hall; 1993.
26. Drover DR, et al. Patient State Index: Titration of delivery and recovery from propofol, alfentanil, and nitrous oxide anesthesia. Anesthesiology 2002;97:82–89.
27. Schultz B, Schultz A, Grouven U. Sleeping stage based systems (Narcotrend). In : Bruch HP. et al., editors. New Aspects of High Technology in Medicine 2000. Bologna: Monduzzi Editore; 2000:285–291.
28. Grouven U, Beger RA, Schultz B, Schultz A. Correlation of Narcotrend Index, entropy measures, and spectral parameters with calculated propofol effect-site concentrations during induction of propofol-remifentanil anesthesia. J Clin Monit Comput 2004;18:231–240.
29. Viertiö-Oja H, et al. Description of the Entropy^{TM} algorithm as applied in the Datex–Ohmeda S/5–Entropy Module. Acta Anaesthesiol Scand 2004;48:154–161.
30. Jensen EW, Lindholm P, Henneberg SW. Autoregressive modeling with exogenous input of middle latency auditory-evoked potentials to measure rapid changes in depth of anesthesia. Methods Inf Med 1996;35:256–260.
31. Litvan H, et al. Comparison of conventional averaged and rapid averaged, autoregressive-based extracted auditory evoked potentials for monitoring the hypnotic level during propofol induction. Anesthesiology 2002;97:351–358.
32. Wong CA, Fragen RJ, Fitzgerald PC, McCarthy RJ. The association between propofol-induced loss of consciousness and the SNAP^{TM} index. Anesthology Analg 2005;100:141–148.
33. Grassberger P, Procaccia I. Measuring the strangeness of strange attractors. Phys D 1983;9:189–208.
34. Widman G, et al. Quantifcation of depth of anesthesia by nonlinear time series analysis of brain electrical activity. Phys Rev E 2000;62:4898–4903.
35. Zhang XS, Roy RJ, Jensen EW. EEG Complexity as a measure of depth of anesthesia for patients. IEEE Trans Biomed Eng Dec 2001;48(12):1424–1433.
36. Bruhn J, Röpcke H, Hoeft A. Approximate entropy as an electroencephalographic measure of anestetic drug effect during desflurane anesthesia. Anesthesiology 2000;92:715–726.
37. Bouillon TW, et al. Pharmacodynamic interaction between propofol and remifentanil regarding hypnosis, tolerance of laryngoscopy, bispectral index, and electroencephalographic approximate entropy. Anesthesiology 2004;100:1353–1372.
38. Shannon CE. A mathematical theory of communication. Bell System Tech J 1948;27:379–423.
39. Bruhn J, et al. Shannon entropy applied to the measurement of the electroencephalographic effects of desflurane. Anesthesiology 2001;95:30–35.
40. Anier A, Lipping T, Melto S, Hovilehto S. Higuchi fractal dimension and spectral entropy as measures of depth of sedation in intensive care unit. Proceedings of the 26-th IEEE EMBS Annual International Conference (EMBC'04), San Francisco; Sept. 2004; pp. 526–529.
41. Lempel A, Ziv J. On the complexity of finite sequences. IEEE Trans Infor Theory 1976;IT–22:75–81.
42. Higuchi T. Approach to an irregular time series on the basis of the fractal theory. Phys D 1998;31:277–283.
43. Jordan KG. Continuous EEG monitoring in the neuroscience intensive care unit and emergence department. J Clin Neurophysiol 1999;16:14–39.
44. Scheuer ML, Wilson SB. Data analysis for continuous EEG monitoring in the ICU: seeing the forest and the trees. J Clin Neurophysiol 2004;21:353–378.
45. Jäntti V, Mustola S, Huotari AM, Koskinen M. The importance of looking at the EEG when presenting uivariate variables to describe it. Br J Anaesth 2002;88:739.
46. Hirsch LJ. Continuous EEG monitoring in the intensive care unit: an overview. J Clin Neurophysiol 2004;21:332–340.
47. Kern SE, X Gie, White JL, Egan TD. Opioid-hypnotic synergy. Anesthesiology 2004;100:1373–1381.
48. Schultz B, et al. Der Narcotrend Monitor: Entwicklung und Interpre-tationsalgorithmus. Anaesthesist 2003;52:1143–1148.

See also ANESTHESIA MACHINES; BLOOD PRESSURE MEASUREMENT; ELECTROENCEPHALOGRAPHY; OXYGEN ANALYZERS; SAFETY PROGRAM, HOSPITAL; TEMPERATURE MONITORING.

MONITORING, AMBULATORY. See AMBULATORY MONITORING.

MONITORING, FETAL. See FETAL MONITORING.

MONITORING, HEMODYNAMIC

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

INTRODUCTION

The word monitor has a variety of meanings, depending on the context. A monitor can be any device for checking on or regulating the performance of a machine, aircraft, or a patient. A patient monitor is usually thought of as something that watches, warns, or cautions if there is a life-threatening event. A more rigorous definition of patient monitoring is Repeated or continuous observations or measurements of the patient, his or her physiological status, and the functions of life support equipment for the purpose of guiding management decisions, including when to make therapeutic interventions and assessment of those interventions (1). As a result, a monitor should not only alert physicians and nurses to potentially life-threatening events, but perhaps should also control devices that maintain life. The primary emphasis of this section deals with hemodynamic monitoring of the critically ill patient who is in an intensive care unit (ICU), but many of the principles apply to all hospitalized patients.

Hemodynamic monitoring relates to monitoring of the blood pressure and blood flow in the cardiovascular system. The cardiovascular system consists of the heart, lungs, and blood vessels, and has a most important function in maintaining life in complex animals, such as humans. Oxygen and fuel must be transported from their source to the individual cells that consume them. The resulting waste products of metabolism must then be disposed of. Thus, the heart and blood vessels transport nutrients to the body and remove the waste products. Clearly, if this system does not function properly, the organism could be compromised. As a consequence clinically applicable methods have been developed to assess the function of the cardiovascular system. Hemodynamic monitoring is one part of this complex monitoring strategy. Typical parameters measured when performing hemodynamic monitoring are heart rate and rhythm, measured through analysis of the electrocardiogram (ECG), blood pressure measurements in various locations in the cardiovascular system, and estimates of blood flow usually using cardiac output as a measure.

THEORY

Hemodynamic monitoring permits minute-to-minute surveillance of the cardiovascular system and provides physiologic data to assist in diagnosis as well as to guide therapy (2–5). The cardiovascular system consists of the heart, lungs, and blood vessels that supply blood to the body and return blood from the peripheral tissue.

It is beyond the scope of this section to describe the detailed anatomy of the cardiovascular system. However, to understand the principles of hemodynamic monitoring knowledge of the functional aspects of the cardiovascular system is essential.

HEART

The heart is made up of four chambers: the right atrium and the right ventricle and the left atrium and the left ventricle (see Fig. 1). The right atrium accepts blood from the systemic circulation (head, arms, and legs) via the superior and inferior vena cava. On atrial contraction the tricuspid valve between the right atrium and right ventricle opens and blood flows into the right ventricle. On ventricular contraction the right ventricle pumps blood through the pulmonic valve into the pulmonary artery and to the lungs where oxygen is added and carbon dioxide is removed. Blood flows from the lungs to the pulmonary veins and then into the left atrium. On atrial contraction the blood flows into the left ventricle through the mitral valve. On ventricular contraction the left ventricle pumps blood through the aortic valve to the systemic circulation (aorta and the systemic vasculature).

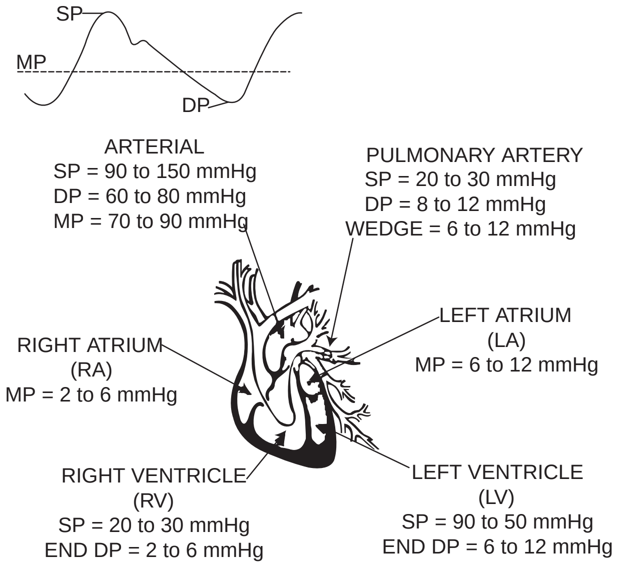


Figure 1. Outline drawing of the heart with its chambers and typical pressures (expressed in mmHg) for each area of the heart. Note the pressures are systolic (SP), diastolic (DP), and mean (MP), as shown on the arterial pressure waveform.

The basic contractile element of the heart is the sarcomere, which is composed of myofilaments, contractile proteins that interdigitate and slide along one another during contraction. Shortening of the sarcomere is the functional unit of heart contraction. Physiologic and pharmacologic agents can change the contractile characteristics of the sarcomeres. Rate and contractility of the heart are controlled by sympathetic and parasympathetic innervation, as well as circulating catecholamines.

Control of Heart Performance

Mechanisms regulating cardiac (heart) output involve not only factors controlling performance of the heart as a pump, but also factors affecting the systemic vascular system and its resistance. Typically, the heart can increase its output to a level of almost five times its resting value. There are two methods by which the heart regulates its cardiac output in response to stress, injury, or disease: by changing heart rate and stroke volume.

Heart Rate Control

Heart rate can be changed rapidly and is thus one of the most effective ways for the heart to change its cardiac output. For a healthy person, an increase in heart rate can more than double the cardiac output when the heart rate increases to near 180 $\text{beats}\cdot\text{min}^{-1}$. However, if a patient with heart disease increases their heart rate to $>120\ \text{beats}\cdot\text{min}^{-1}$ they may have deleterious responses because of the increased demand for oxygen by the heart muscle. Blood flow in the heart muscle occurs primarily during diastole (the relaxation phase of heart contraction). Increasing heart rate decreases the time for cardiac circulation during diastole. In normal subjects, decreasing the heart rate to $\sim$ 50 beats$\cdot$min^{-1} may not decrease cardiac output because there is increased diastolic filling time that increases stroke volume.

Stroke Volume Changes

The stroke volume of an intact ventricle is influenced by (*1*) ventricular end-diastolic volume (called preload), (*2*) ventricular afterload, and (*3*) contractility.

Preload. Preload is the term used to define the end-diastolic stress in the wall of the ventricle. For example, zero preload would result in the ventricle ejecting no blood. However, with increased preload, ventricular ejection generally increases linearly until the capacity of the pump (heart) is exceeded. Since the end-diastolic volume so profoundly influences the myocardial fiber length it has a great influence on the myocardial performance. The Frank–Starling law describes this principle and is illustrated graphically in Fig. 2. The most accessible measure of right ventricular preload is the right atrial pressure. Left atrial pressure is used to estimate left ventricular preload. Since the left ventricle does most of the work of the heart, it is usually the first part of the heart muscle to fail. Consequently, the measurement or estimation of the left atrial pressure is important in assessing a patient's hemodynamic status.

Afterload. Afterload is a measure of the impedance (resistance) against which the right or left ventricles must eject blood. Resistance (R) is calculated by measuring blood flow and pressure and then using Ohm's law {Eq. 1}.

$$R = \frac{\text{mean blood pressure}}{\text{cardiac output}} \tag{1}$$

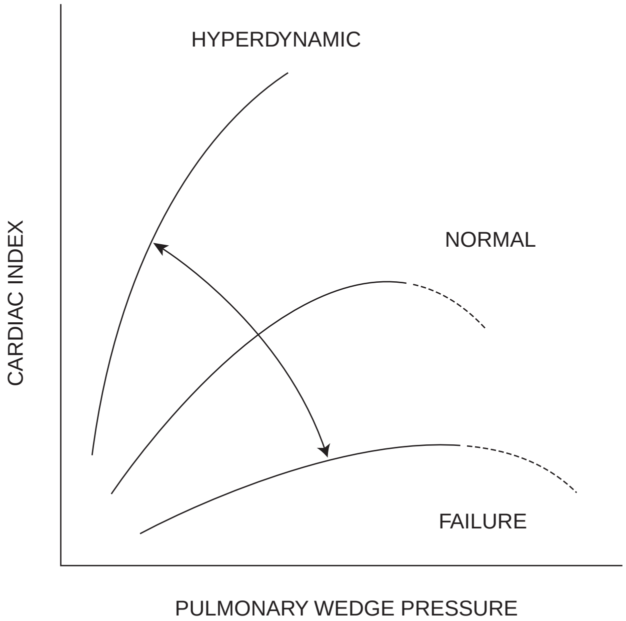


Figure 2. Frank–Starling curve of the heart showing the ventricular performance (cardiac index) plotted against the end-diastolic volume typically estimated by using pulmonary artery wedge pressure. Note to the right of these curves, there is a pulmonary wedge pressure above which the heart in ineffective in producing increased flow.

Systemic Circulation

Blood flow to the periphery of the body is controlled by local autoregulation and by the autonomic nervous system. Local autoregulation of blood flow helps tissue meet its oxygen requirements. For example, with decreased blood flow, metabolic byproducts increase, causing local vasodilatation that tends to increase blood flow. There are baroreceptors, similar to blood pressure transducers, located in the aortic arch and the carotid sinus which sense blood pressure. Via the baroreceptor reflex mechanism, the body regulates the blood pressure. In addition, chemoreceptors in the carotid sinus and other locations regulate respiration by responding to changes in CO_2 and O_2.

Pulmonary Circulation

The pulmonary arterial vessels differ markedly from systemic arterial vessels; they have thinner walls, less muscle, and have a resistance to blood flow about one-sixth that of the systemic circulation.

Contractility. Contractility is a measure of how a healthy heart performs. A healthy heart pumps vigorously and nearly empties its ventricles with each beat and is said to have excellent contractility. On the other hand, a compromised heart may not be able to empty effectively.

HEMODYNAMIC MONITORING

Bedside hemodynamic monitoring makes use of data gathering procedures that were formerly only done in diagnostic cardiac catheterization laboratories. Understanding the relationship between the pressure and blood flow in the cardiovascular system is the primary reason for performing hemodynamic monitoring. The cardiovascular system responds to many and varied stimuli and can be affected by physical conditioning, drugs, disease, blood loss, injury, and myocardial insult such as a heart attack. Because of the complexity of factors controlling the body, it is necessary to make hemodynamic measurements on the system to understand disease processes and provide optimum therapy to the patient.

Electrocardiogram

Electrocardiogram (ECG) monitoring is used to determine heart rate and detect arrhythmias, pacemaker function, and myocardial ischemia (lack of blood flow to the heart muscle). To permit optimum ECG monitoring the signal quality must be excellent (6). Since the ECG electrical signal from the heart is only 0.5–2.0 mV at the skin's surface, it is best measured by properly preparing the skin and optimally placing the electrodes. Skin can be properly prepared by removing oil from the surface and abrading the skin to remove the dry dead outer layer (stratum granulosum). In 90% of patients, proper skin preparation reduces electrode resistance from as high as 200 to as low as 10 kΩ. Good electrode placement allows the electrodes to receive the maximum ECG signal with minimum noise. By placing the electrodes over bony prominence, such as the sternum or clavicles, muscle artifact (EMG) can be reduced. Motion artifact caused by movement of electrodes can be minimized not only by proper skin preparation, but also by taping a strain-relieving loop in the lead wires to prevent movement artifact. Shielded wire on the ECG leads helps minimizes pickup of alternating current (ac) electrical fields from 60 Hz power sources, electrosurgical units, and other sources like radio transmitters. The two leads that connect the patient form a loop through which magnetic fields pass and can induce unwanted voltages. Pickup from magnetic fields can be minimized by decreasing the loop area, by keeping the lead wires close together (usually twisted pairs), and by avoiding draping the lead wires over motors, lights, or other electrically powered instruments.

Electrocardiogram Arrhythmia Monitoring

Early in the development of monitoring techniques, the application of computer technology to detect patterns of the electrocardiogram caught the attention of those who sought to improve care of the critically ill. The computer appeared to be a logical candidate for relieving the nursing and medical staff of the tedious chore of continuously visually monitoring a multichannel oscilloscope.

Arrhythmia monitoring is one of the most sophisticated of the bedside monitor's tasks. People-based arrhythmia monitoring is expensive and unreliable, and those who do it find the task to be tedious and stressful. Today virtually every bedside monitor has rhythm monitoring built in. These monitors use computers and a variety of algorithms to detect and classify ECG rhythm abnormalities. Classifying these rhythm abnormalities is important to hemodynamic monitoring since irregular rhythms can cause dramatic inefficiencies in how the heart works as a pump. For example, Fig. 3 shows three strip recordings of the ECG and the corresponding pressure waveform from three different arrhythmias (ventricular tachycardia, couplet, which is two beats of abnormal electrical origin, and bigeminy where every other beat is from abnormal electrical origin). Note that several of the abnormal beats are hardly effective at creating any change in the arterial pressure. Those same beats deliver small stroke volumes to the patient's systemic circulation. As a consequence, one cannot assume that the cardiac output remains constant or increases just because the heart rate increases.

MEASUREMENTS

Blood Pressure Monitoring

Arterial blood pressure can be measured by both direct and indirect methods. However, central venous pressure (CVP), pulmonary artery (PA), and pulmonary capillary wedge pressure (PCWP) used to estimate left atrial pressure, at present, can only be measured by direct invasive methods.

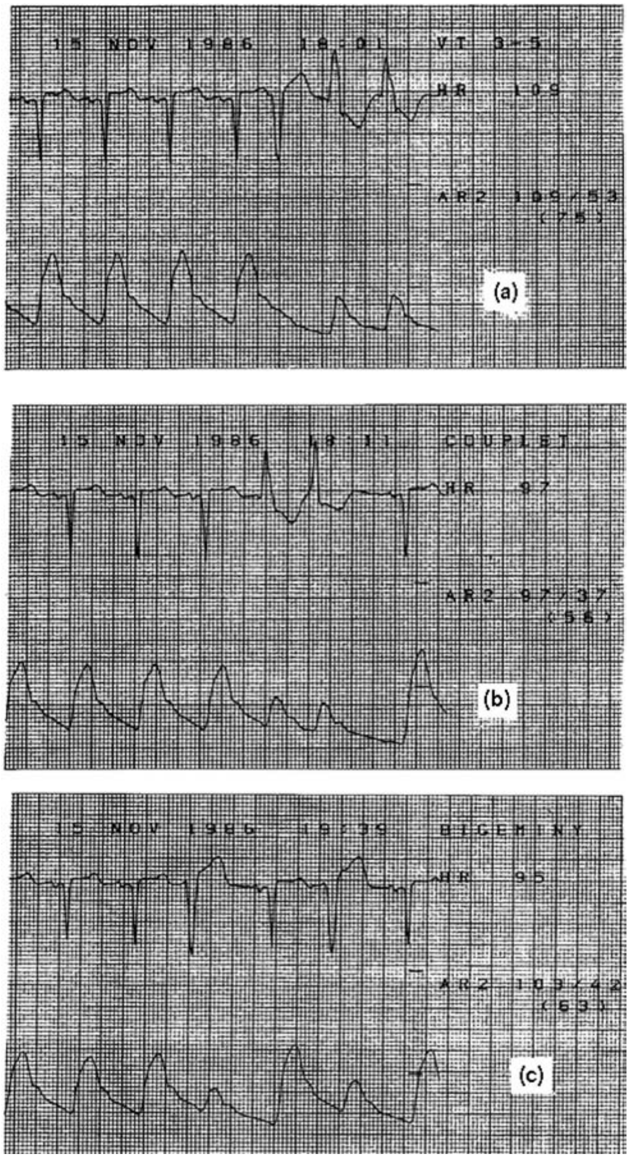


Figure 3. Electrocardiogram and arterial pressure waveforms with three different abnormal rhythms. (a) Ventricular tachycardia (VT), which occurs during the last two beats of the strip. (b) Couplets where two successive beats have an abnormal electrical origin. (c) Bigeminy where every other beat is from an abnormal electrical origin. Pressures are expressed in millimeters of mercury. For example, the patient in (a) has a systolic arterial pressure of 109 mmHg, a diastolic pressure of 53 mmHg, and a mean arterial pressure of 75 mmHg.

Arterial Blood Pressure: Indirect Measurement; Using a Cuff. Recently, the American Heart Association has updated its recommendations on accurate "indirect" measures of blood pressure (7). The update reports that the auscultatory technique with trained observer and mercury manometer continues to be the method of choice for measurement of a patient's blood pressure in a physician's office. The report also suggests that hybrid devices that use electronic transducers rather than mercury have promise. The report indicates that oscillometric devices can also be used, but only after careful validation.

Unfortunately, the indirect measurement of arterial pressure has serious limitations for patients in shock usually signaled by low blood pressure. Also, since virtually all reliable indirect pressure measurement techniques require cuff inflation, such measurements can only be made intermittently.

Direct Blood Pressure Measurements. The direct measurement of blood pressure allows for continuous and accurate assessment of blood pressures. Direct and continuous pressure monitoring allows detection of dangerous hemodynamic events and provides the information necessary to initiate and regulate patient therapy to prevent catastrophic events. However, monitoring of pressures provides valuable information only when it is obtained in a technically satisfactory manner.

To accomplish direct blood pressure measurements, it is necessary to insert a catheter directly into the cardiovascular system (8). This invasive technique has risks that must be weighted against the benefits that can be obtained. These risks include, infection, blood loss, insertion site damage and other factors (9,10). For many patients who are in shock or who have cardiac disease, the benefits far outweigh the risks. Formal methods for assessing these risks have been published by the Coalition for Critical Care Excellence (11).

Blood pressure can be measured on both the pulmonary (right heart) and systemic (left heart) sides of the circulatory system. Measurements of both pulmonary and systemic parameters yield different and important cardiovascular status. The CVP reflects the patient's circulating blood volume or venous tone, and right atrial and ventricular pressures (right ventricular preload). To measure the right atrial pressure accurately a catheter must be placed in a major vein within the chest or directly in the right atrium. The CVP values fluctuate about atmospheric pressure. The level of the right heart is usually taken as the zero reference point from which all other blood pressures are measured. The CVP gives an indication of only the function of the right heart, and not left heart's performance.

To measure the left atrial pressure, it is necessary to place a catheter tip through the atrial septum from the right atrium (usually done only with fluoroscopic control in the cardiac catheterization laboratory) or estimating it by placing a pulmonary artery (Swan–Ganz) catheter in the wedged position by inflating its balloon near the catheter tip.

EQUIPMENT

Components of Direct Pressure Monitoring Systems

The components of a direct blood pressure monitoring system for critically ill patients are shown in Fig. 4 (6,8). The components numbered 1–7 in the figure are known as

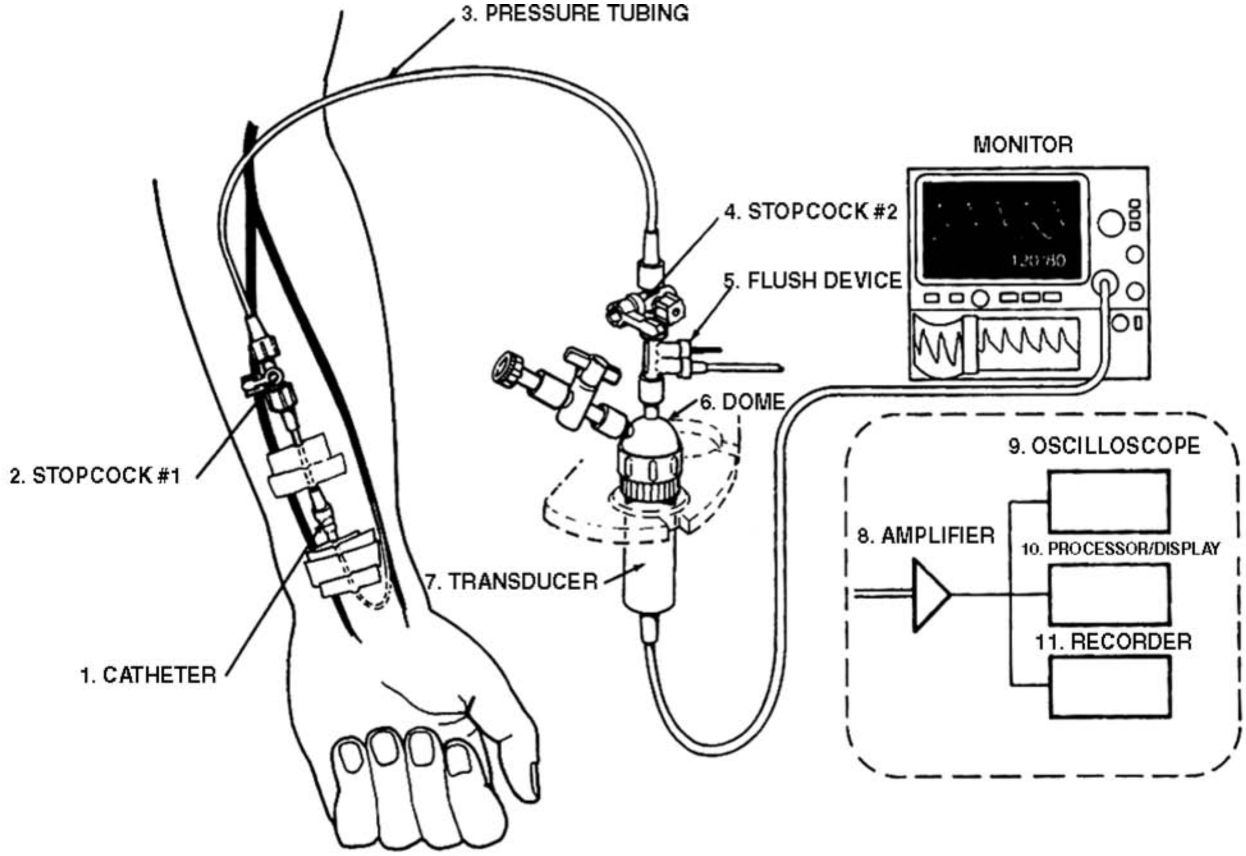


Figure 4. The 10 components used to monitor direct blood pressure. The monitoring components are nearly independent of whether the catheter is in an artery (radial, brachial, or femoral) or in the pulmonary artery. Size of transducer and plumbing components are enlarged for illustration purposes. [Reproduced from Ref. 6, with permission.]

the "plumbing" system and must always be sterile because the fluid contained therein comes in direct contact with the patient's blood. Today virtually all of these components are disposable or single-use items to minimize patient infection. Components 8–11 in Fig. 4 are used for processing and displaying pressure waveforms and derived hemodynamic parameters.

1. Catheter. Arterial and pulmonary artery catheters provide access to the patient's blood vessels to (a) monitor intravascular pressure and (b) provide a site for samples for blood to allow determination of blood gas and other laboratory testing parameters. These catheters are typically placed by the percutaneous method, either by the Seldinger "over-the-needle" technique or by introducing the catheter through a needle (8).
2. Sampling stopcock. Stopcock 1 is used as a site for withdrawing blood for analysis. When filling the catheter-tubing-transducer system with fluid, precautions must be taken to be sure all central switching cavities of the stopcock are filled and that entrapped air bubbles are removed. Because stopcocks are especially vulnerable sources of patient contamination, these devices must be handled with extreme care; ports not in active use should be covered with sterile caps and medical personnel should never touch open ports of the stopcocks.
3. Pressure tubing. The catheter and stopcock are normally attached to a continuous flush device and transducer by noncompliant pressure tubing. To optimize the dynamic response of the catheter-tubing-transducer system, long lengths of tubing must be avoided.
4. Stopcock 2. This stopcock is usually put in place to allow disconnection of the flush device and transducer from the patient when the patient is moved or when initially filling the system with fluid.
5. Continuous flush device. This device is used not only when initially filling the pressure monitoring system, but also to help prevent blood from clotting in the catheter. These devices provide a "continuously flush" of fluid at a rate of from 1 to 3 $\text{mL} \cdot \text{h}^{-1}$.

6,7. Transducer dome and Pressure transducer. Today virtually all transducers used for monitoring are highly reliable, standardized, disposable devices (12,13).

8. Amplifier system. The output voltage required to drive an oscilloscope or strip-chart recorder is provided by an amplifier system inserted between the transducer and display. Pressure transducer excitation is provided either from a direct current (dc) or ac source at a voltage of 4–8 V revolutions per second (rms). Most amplifier systems include low pass filters that filter out unwanted high frequency signals. Pressure amplifier frequency response should be flat from 0 to 50 Hz to avoid pressure waveform distortion.
9. Oscilloscope. Pressure waveforms are best visualized on a calibrated oscilloscope or other similar display panel.
10. Digital processing and display. Digital displays provide a simple method for presenting quantitative data from the pressure waveform. They are found on most modern pressure monitoring equipment. Systolic, diastolic, and mean pressure are typically derived from the pressure waveforms.
11. Strip-chart recorders. Frequently, strip-chart recorders are used to document dynamic response characteristics, respiratory variations in pulmonary artery pressures, and aberrant rhythms and pressure waveforms.

STATIC CALIBRATION

Zeroing and calibrating the transducer are two important steps in setting up the direct pressure-monitoring system.

Zeroing the Transducer

The accuracy of blood pressure readings depends on establishing an accurate reference point from which all subsequent measurements are made. The patient's midaxillary line (right heart level) is the reference point most commonly used (14). The zeroing process is used to compensate for offset caused by hydrostatic pressure differences, offset in the pressure transducer, amplifier, oscilloscope, recorder, and digital displays. Zeroing is accomplished by opening an appropriate stopcock to the atmosphere and aligning the resulting fluid-air interface with the midaxillary reference point.

Once the system is zeroed the stopcock can be switched to allow the patient's waveform to be displayed. Pulmonary artery and pulmonary artery wedge pressure are especially susceptible to improper zeroing and should be measured only after the zero point has been verified.

Sensitivity Calibration

The sensitivity of most pressure transducers is fixed at $5.0\ \mu\text{V} \cdot \text{V}^{-1}$ of excitation applied per 1 mmHg (0.13 kpa) and calibrated by the manufacturers to within ±1%. This degree of accuracy is adequate for clinical purposes. As a consequence standardized transducers need only to be zeroed to obtain accurate pressure measurements (12,13).

CHECKING DYNAMIC RESPONSE

In the critical care setting, where most hemodynamic monitoring is carried out, the catheter-tubing-transducer systems used can usually be characterized as an underdamped second-order dynamic system analogous, for example, to a bouncing tennis ball. A second-order dynamic system can be expressed mathematically by a second-order differential equation with characteristics determined by three mechanical parameters: elasticity, mass, and friction. These same parameters apply to a catheter-tubing-transducer system where the natural frequency (f_n in Hz) and damping coefficient determine the dynamic characteristics for a catheter-tubing-transducer system. For an underdamped second-order system f_n and define the system's

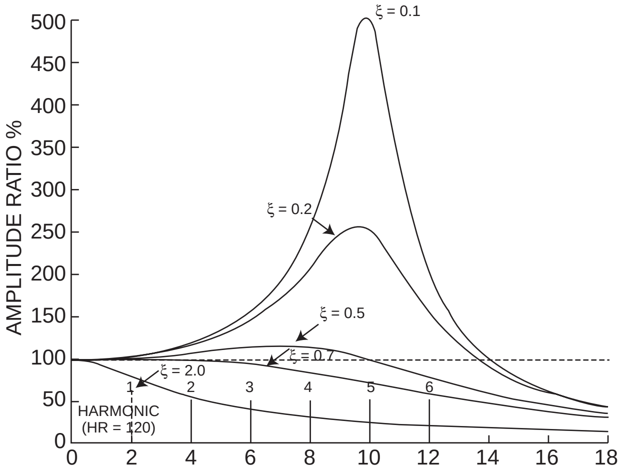


Figure 5. Family of frequency versus amplitude ratio plots for five different damping coefficients ζ and natural frequencies f_n of the plot shown is 10 Hz. When $\zeta = 0.1$, the system is very underdamped, and when $\zeta = 2.0$, it is overdamped. The dashed line shows the frequency versus amplitude characteristic that would occur if the system had a flat frequency response. Along the frequency axis are plotted the harmonics of the pressure wave if the heart rate were 120 beats $\cdot$ min^{-1} (2 beats $\cdot$ s^{-1}). Note that by the fifth harmonic (10 Hz) if $\zeta = 0.1$ the true signal would be amplified five times. If $\zeta = 2.0$ there would be an attenuation to about one-fourth of the amplitude. In both cases there would be a gross waveform distortion because neither situation reflects a high fidelity system dynamic response. Fidelity of the system can be improved by increasing the f_n or adjusting ζ to be in the range of 0.5–0.7. [Reproduced from Ref. 6, with permission.]

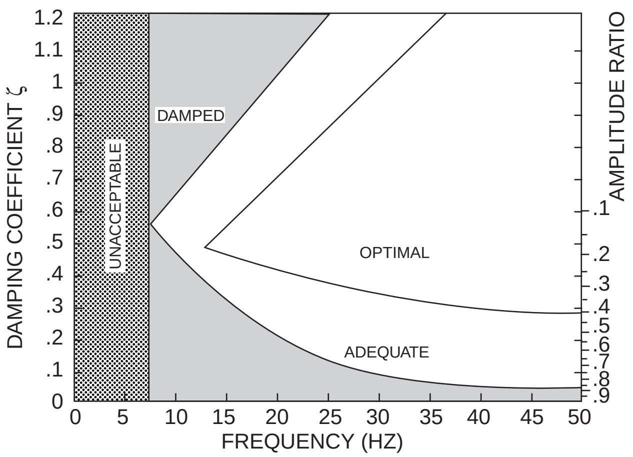


Figure 6. Plot of f_n versus ζ illustrating the five areas into which catheter-tubing-transducer systems fall. Systems in the optimal area will reproduce even the most demanding (fast heart rate and rapid systolic upstroke) arterial or pulmonary artery waveforms without distortion. Systems in the adequate area will reproduce most typical patient waveforms with little or no distortion. All other areas will cause serious and clinically important waveform distortion. Note the scale on the right can be used to estimate ζ from the amplitude ratio determined during fast flush testing (11). See Fig. 8 for an example of waveforms. [Reproduced from Ref. 6, with permission.]

dynamic characteristics. In the clinical setting f_n and can be measured easily and conveniently by using the "fast-flush" method.

Dynamic response characteristics of catheter-tubing-transducer systems have been defined by two interrelated techniques. The first technique specifies the frequency bandwidth and requires that the system frequency response must be flat up to a given frequency so that a specified number of harmonics usually 10 of the original pulse wave can be reproduced without distortion (Fig. 5).

The second technique specifies f_n and The plot of f_n and in Fig. 6 has five areas (6,15). If the characteristics of the catheter-tubing-transducer "plumbing" system fall in the adequate or optimal area of the graph, the pressure waveforms will be adequately reproduced. If the characteristics fall into one of the remaining three areas, there will be pressure waveform distortion. Most catheter-tubing-transducer systems assembled under optimal conditions are underdamped, but a few fall into the unacceptable areas of the chart. Methods for optimizing the catheter-tubing-transducer system components have been outlined (15–20). In the clinical setting, there are dramatic differences between each patient setup; therefore it is mandatory to verify the adequacy of each pressure-monitoring system by testing them. The testing can be done easily using the fast-flush technique.

A fast flush is produced by opening the valve of the continuous flush device, for example, by pulling and quickly releasing the pigtail valve on a continuous flush device. The rapid valve closure generates a near square wave pressure signal from which f_n and of the catheter-tubing-transducer system can be measured.

Once the fast-flush test has been executed two or three times, the dynamic response characteristics (f_n and) can quickly and easily be determined. Natural frequency f_n can be estimated by measuring the period of each full oscillation on a strip-chart recorder following a fast flush (Fig. 7a) and calculating the frequency from the period. Damping coefficient can be determined by using the amplitudes of any two successive peak-to-peak values measured after a fast flush. The amplitude ratio is calculated by dividing the measured height of the smaller peak-to-peak value by that of the amplitude of the larger peak-to-peak value (Fig. 7b). The amplitude ration can then be converted to a damping coefficient by using the scale in the right side of Fig. 6.

Once f_n and have been determined, these data can be plotted on the graph of Fig. 6 to ascertain the adequacy of dynamic response. Some bedside monitors and recorders may compromise the clinical user's ability to use the fast-flush technique because the monitors have built-in low-pass filters. These filters should be expanded to at least 50 Hz or be eliminated.

Several factors lead to poor dynamic responses: (*1*) air bubbles in the system usually caused by a poor initial catheter-tubing-transducer system setup, (*2*) pressure tubing that is too long, too compliant, or a diameter that is too small, and (*3*) pressure transducers that are too compliant. The best way to enhance the system's dynamics is to improve f_n.

Invasive pressure monitoring systems have patient risks, such as a source of infection and air embolism. In addition, great care is required by clinical users to optimize

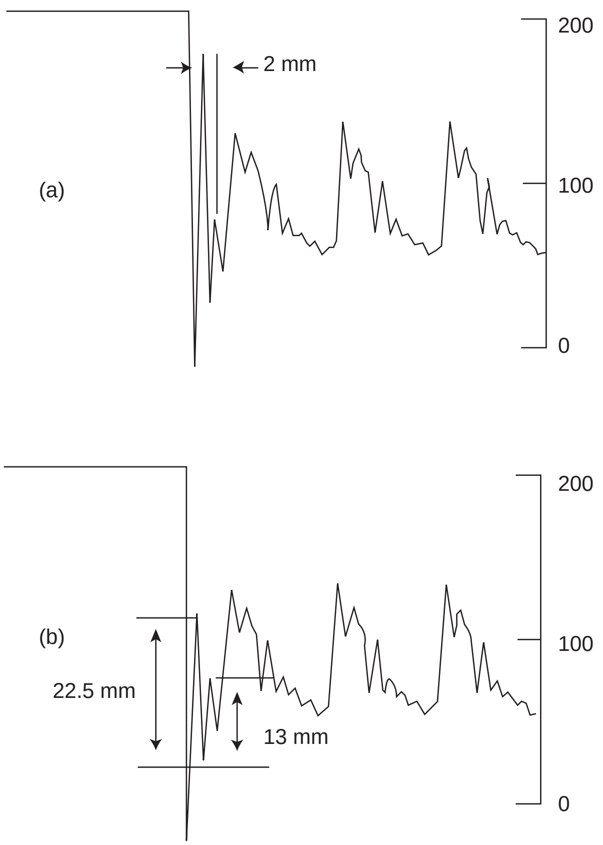


Figure 7. Measuring dynamic response parameters from the fast-flush waveform, (a) The natural frequency f_{n} can be determined by using a strip-chart recording to measure the period of one full oscillation, as shown. In this example, one full cycle is 2 mm and at a paper speed of 25 mm $\cdot$ s^{-1} this results in $f_{\mathrm{n}} = 12.5$ Hz $=$ 25 mm $\cdot$ s^{-1}/2 mm. (b) Determining the damping coefficient ζ required measuring two successive peak-to-peak values of the resulting oscillations. The amplitude ratio of the two successive peaks is taken, giving a value of $0.58 = 13/22.5$. With use of the amplitude ratio and the scale on the right side of Fig. 6, the damping coefficient $\zeta = 0.17$. Plotting the natural frequency and damping coefficient on Fig. 6 shows that this system is underdamped.

dynamic response and proper zeroing to provide accurate and reliable data. Merely looking at pressure waveforms will not provide the information required to determine the adequacy of the system's dynamic response (see Fig. 8). Fast-flush testing to determine these parameters is essential.

SIGNAL AMPLIFICATION, PROCESSING, AND DISPLAY

Once the pressure signal has been transmitted to the transducer, the bedside monitor operates on that signal. Most monitors not only display the heart rate and systolic, diastolic, and mean pressure, but they also display the processed waveform on an oscilloscope and provide an analog output for a recorder or for transmission to a central display.

Placement of the Pulmonary Artery Catheter

The balloon-tipped, flow-directed, pulmonary artery catheter (Swan–Ganz) came into widespread use in 1970 (21). The follow-up development by Ganz of a practical thermal dilution attachment to the pulmonary artery catheter permitted convenient and easy measurement of cardiac output (22). Since these early developments with the Swan–Ganz catheter, the pulmonary artery catheter has been fitted with optical fibers which allow measurement of mixed venous oxygen saturation (23).

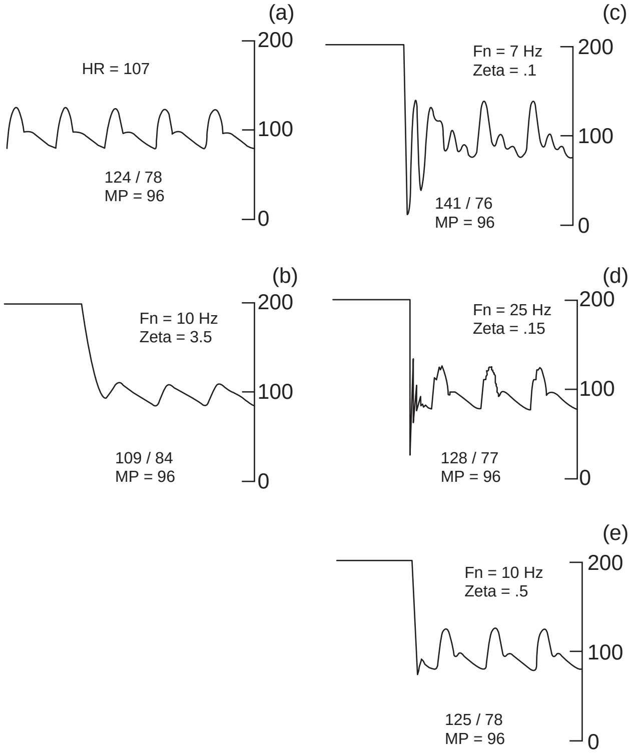


Figure 8. Arterial pressure waveforms were obtained from the same patient. Shown are Systolic/Diastolic and Mean Pressure (*MP*). In panel the (a) Patient's actual arterial pressure waveform as if recorded with a catheter-tipped transducer is shown, (b) shown the same patient's arterial waveform recorded with an overdamped system ($\zeta = 3.5$). Note the fast-flush signal (upper left) returns slowly to the patient waveform. Systolic pressure is underestimated, diastolic pressure is overestimated, and *MP* is unchanged, (c) An underdamped condition ($\zeta = 0.1$) with low $f_n = 7$ Hz. After the fast flush, the pressure signal oscillates rapidly (rings). Systolic pressure is overestimated, diastolic is slightly underestimated, and MP is correct, (d) shows an underdamped condition ($\zeta = 0.15$), but with high $f_n = 25$ Hz. The pressure waveform is slightly distorted and systolic, diastolic, and mean pressures are close to the actual pressures, (e) shown an ideally damped pressure monitoring system ($\zeta = 0.5$). The undershoot after the fast flush is small and the original patient waveform is adequately reproduced. [Reproduced from Ref. 6, with permission.]

The pulmonary artery catheter is inserted into the right side of the circulation using the percutaneous technique typically using entry from either the internal jugular or the subclavian vein. The catheter is floated into the pulmonary artery without use of fluoroscopy, using the hemodynamic pressure waveforms as a guide (Fig. 9).

Accurate Measurement of Pulmonary Artery Pressure. Since it was introduced, the balloon-tipped, flow-directed, pulmonary artery catheter (Swan–Ganz) has been widely used in intensive care units. The ease with which it is

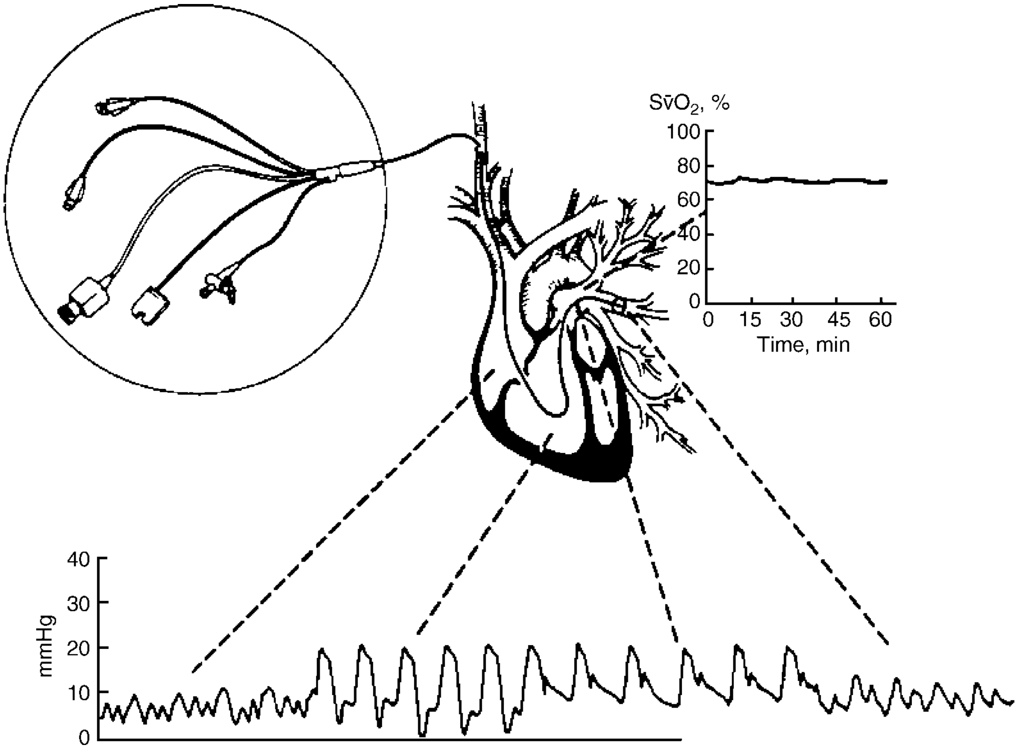


Figure 9. Composite illustration showing normal pressure waveforms obtained as a fiber optic balloon flotation pulmonary artery catheter (Swan-Ganz) is advanced from the right atrium to the pulmonary artery wedge position. [From Daily and Tilkian in Reading List (1986), with permission.]

usually inserted may lead one to conclude that the measurements of pulmonary artery and wedge pressure (PCWP) are easily and reliably measured. However, such is not the case.

Pulmonary artery pressures can be measured accurately only if the following steps are taken (24–27):

1. The monitor is properly zeroed.
2. Strip-chart recordings of all PA pressures for a time period covering at least three respiratory cycles are obtained. Using only the monitor's digital displays is insufficient.
3. Dynamic response testing (fast flush) should be obtained when the catheter is in each position (i.e., wedge and PA). If the dynamic response is not adequate, the problems with the catheter-tubing-transducer system must be resolved before accurate pressures can be measured.
4. Pressures (i.e., systolic, diastolic, and mean pressures) should be assessed from a monitor's display or a strip-chart recording. The pressure measures should be made at the end expiration when the transmural pressure is nearest zero.

CARDIAC OUTPUT DETERMINATION

Cardiac output is the volume of blood ejected by the heart every minute. Cardiac output is a helpful measurement since it can be used to evaluate the overall cardiac status of the critically ill patient, as well as help make the diagnosis of cardiovascular disease. Ideally a cardiac output measurement system would be continuous, automatic, minimally invasive, accurate, fast, inexpensive, and easy to use clinically. The most common method used to measure cardiac output in critically ill patients is still the indicator dilution method. The pulmonary artery catheter (Swan–Ganz) introduced in the 1970s revolutionized the ease with which cardiac output could be measured.

The thermal dilution method requires injection of cold physiological solution, usually normal saline, into the superior vena cava or right atrium. Cardiac output is determined by measuring the area under the time–temperature curve measured in the pulmonary artery that results from the injection of the cold solution.

The thermal dilution method for determining cardiac output relies on several assumptions that are not always correct. First, the exact amount of thermal indicator injected cannot be quantitated precisely. Second, indicator is lost at various stages and this loss of indicator (heat loss) leads to errors.

A block diagram of the thermal dilution measuring system with typical thermal dilution curves and time of injection indicated are shown in Fig. 10. Figure 10c and d show the transit time for the cooled blood moving from the injection site in the right atrium to the pulmonary artery measurement site. Calculation of cardiac output requires measuring the area under the curve. Consequently, a baseline temperature must be established before the injection. In turn, the end point is usually determined by extrapolating to the baseline temperature.

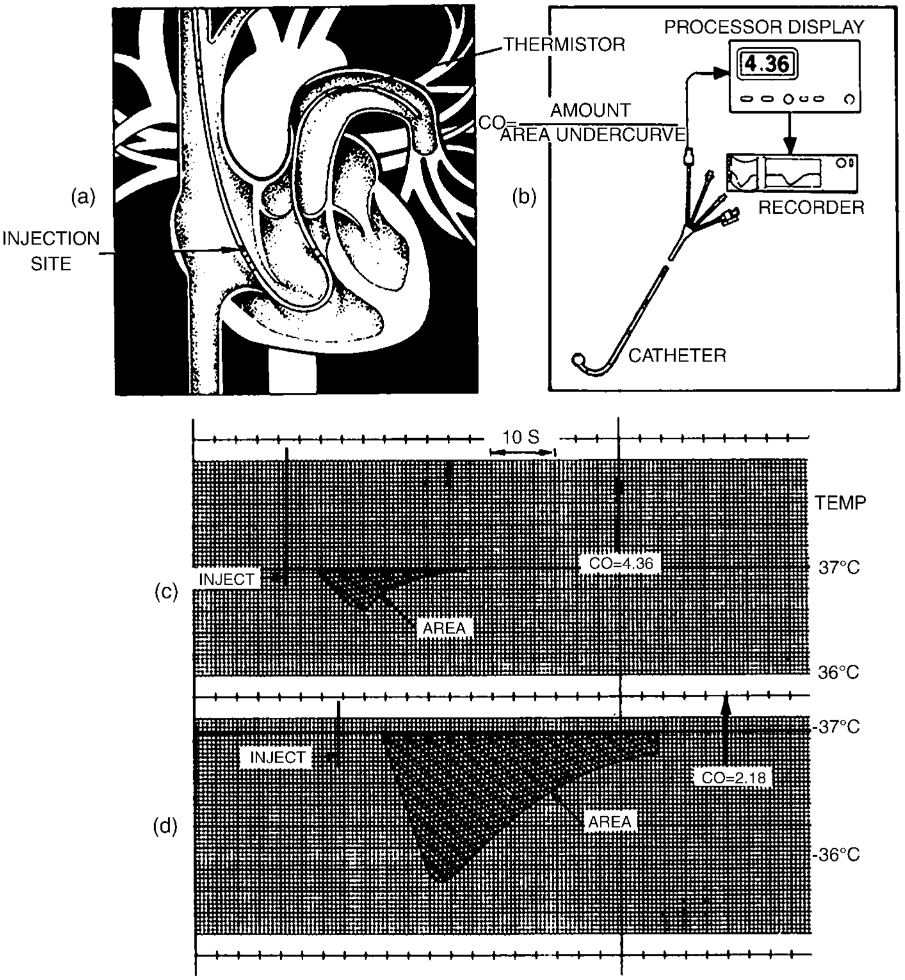


Figure 10. Schematic diagram of the thermal dilution measurement of cardiac output. A recorder of some type should always be used to verify the quality of the thermal dilution curve, (a) shows the thermal dilution catheter placed into the pulmonary artery. Note the location of injection site and thermistor, (b) shows the connection of the thermal dilution catheter connected to a cardiac output processor and recorder, (c) shows a typical temperature-time plot sensed by the thermistor near the catheter tip after an iced saline injection. The cardiac output determined in this case was 4.36 $L \cdot min^{-1}$. (d) Shows a similar temperature–time plot for a patient with low cardiac output (2.18 $L \cdot min^{-1}$). Note the larger area and broader dispersion of the waveform caused by the lower flow. [Reproduced from Ref. 9, with permission.]

To ensure that accurate thermal dilution cardiac output results are obtained, it is recommended that the thermal dilution curves be presented on a monitoring screen or on a strip-chart recorder. Studies have shown that synchronizing the injections with the respiratory cycle improves the technique's reproducibility (28). Since there is considerable variability in cardiac output between measures, at least three reproducible curves are usually obtained. Averaging the findings from these three curves gives a more representative assessment of cardiac output.

In recent times, the complexities of using the Swan–Ganz pulmonary artery catheter have result in controversies. Some clinicians feel that such systems should only be used only when needed and then only sparingly while others have a differing viewpoint (29,30). Still others have questioned the ability of making accurate central venous and pulmonary artery occlusion pressures and whether it matters (26,27). Many of those issues will be resolved in the future when there might be better methods for measuring hemodynamic parameters. Until that time, physicians and nurses caring for critically ill patients who require hemodynamic monitoring should be aware of several how to oriented publications (31–34).

Alarming Based on Hemodynamic Parameters. Clinical hemodynamic monitoring is now several decades old. What started from a simple beginning has since seen many dramatic changes in both the development of new medical devices and skills of the clinicians to use those devices. However, it is my feeling that we are not yet at optimum hemodynamic monitoring. Some recent publications on the topic are illustrative. Sander and colleagues in Germany have recently looked at categories of patients with elevated heart rates who are at higher risk of cardiac complications (35). Their work resulted in an editorial comment Vital are vital signs (36). Additional recent work at Vanderbilt University indicates that volatility in vital signs is a good predictor of death in trauma patients (37). Finally, the problem of false alarms continues to be a huge problem with current bedside monitors. As part of a Master of Science thesis in Medical Informatics at the University of Utah, an investigator found that only about one-third of the standard alarms for patients in a variety of ICU care were true alarms. Thus about two-thirds of the alarms are false. However, if the alarming system used heart rates determined from both the ECG and the Arterial Blood Pressure, the number of alarms decreased by ~ 50% and the false alarm rate was only ~ 25% (38). Having smarter and better hemodynamic monitoring with better and smarter alarming systems will be crucial for to future monitoring systems.

COMPUTERIZED DECISION SUPPORT

Much has been learned about hemodynamic measurements and how to use the data to calculate derived patient parameters. These parameters can then be used to determine patient status and augment patient therapy (39–42).

Using hemodynamic data available from bedside monitors and combining that data, in a structured and coded electronic patient record allows for optimal computerized decision support (42–44). Morris and his colleagues have stated the value of computerized decision support well (45). Only adequately explicit protocols contain enough detail to lead different clinicians to the same decision when faced with the same clinical scenario. Guidelines of care provide only general guidance to patient care and require clinicians considerable latitude in which care decision should be made. Computerized protocols, on the other hand, can be patient-specific and evidence based (46). Using computerized decision-support tools variation in clinical practice can be reduced and favorable effects on improve patient outcomes can be accomplished (45,46).

FUTURE

There are still needs for improvement in hemodynamic monitoring. Being able to make the measurements continuously, less invasively, and more reliably are areas where progress is needed. Clearly, using computer aided decision-support technology to help reduce false alarms and to guide clinicians in making better patient diagnosis and more timely and more optimal and effective treatment decision offer ample opportunity for future research and progress.

BIBLIOGRAPHY

1. Gravenstein JS, Paulus DA. Monitoring Practice in Clinical Anesthesia. Philadelphia (PA): Lippincott; 1982.
2. Bruner JMR. Handbook of Blood Pressure Monitoring. Littleton (MA): PSG Publishing Co.; 1978.
3. Daily EK, Schroeder JS. Techniques in Bedside Hemodynamic Monitoring. 3rd ed. St. Louis (MO): Mosby; 1985.
4. Pinsky MF. Functional hemodynamic monitoring. Intensive Care Med 2002;28:386–388.
5. Pinsky MF. Hemodynamic monitoring in the intensive care unit. Clin Chest Med 2003;24:549–560.
6. Gardner RM, Hollingsworth KW. Optimizing ECG and pressure monitoring. Crit Care Med 1986;14:651–658.
7. Pickering TG, et al. Recommendations for blood pressure measurement in humans and experimental animals: part 1: Blood pressure measurement in humans: A statement for professionals from the subcommittee of professional and public education of the American Heart Association council on high blood pressure research. Circulation 2005;111:697–716.
8. Gardner RM. Hemodynamic monitoring: From catheter to display. Acute Care 1986;12:3–33.
9. Gardner RM, Schwartz R, Wong HC, Burke JP. Percutaneous indwelling radial-artery catheters for monitoring cardiovascular function (Prospective Study of the Risk of Thrombosis and Infection). N Engl J Med 1974;290:1227–1231.
10. Kline AM. Pediatric catheter-related bloodstream infections: Latest strategies to decrease risk. AACN Clin Issues 2005;16: 185–198.
11. Bone RC, et al. Standards of evidence for the safety and effectiveness of critical care monitoring devices and related interventions. Coalition for critical care excellence: Consensus Conference on Physiological Monitoring Devices. Crit Care Med 1995;23:1756–1763.
12. Kutik MH, Gardner RM. Standard for Interchangeability and Performance of Resistive Bridge Blood Pressure Transducers, Arlington (VA): Association for the Advancement of Medical Instrumentation (AAMI); 1986.
13. Gardner RM. Accuracy and reliability of disposable pressure transducers coupled with modern pressure monitors. Crit Care Med 1996;24:879–882.
14. McCann II UG, et al. Invasive arterial BP monitoring in trauma and critical care. Effects of variable transducer level, catheter access, and patient position. Chest 2001; 120:1322–1326.
15. Gardner RM. Direct blood pressure measurement dynamic response requirements. Anesthesiology 1981;54(3):227–236.
16. Gardner RM. Blood pressure monitoring. In: Webb AR, Shapiro MJ, Singer M, Suter PM, editors. Oxford Textbook of Critical Care. Oxford University Press. 1998; p 1087–1090. chapt 16.
17. Gardner RM. Fidelity of recording: Improving the signal-tonoise ratio. In: Martin J. Tobin, editors. Principles and Practice of Intensive Care Monitoring. New York: McGraw-Hill; 1997. p. 123–132. chapt. 8.
18. Kleinman B, Powell S, Kumar P, Gardner RM. The fast flush test measures the dynamic response of the entire blood pressure monitoring system. Anesthesiology 1992;77:1215–1220.
19. Kleinman B, Powell S, Gardner RM. Equivalence of fast flush and square wave testing of blood pressure monitoring systems. J Clin Monit 1996;12(2):149–154.
20. Promonet C, et al. Time-dependent pressure distortion in a catheter-transducer system: Correction by fast flush. Anesthesiology 2000;92:208–218.
21. Swan HJC, et al. Catheterization of the heart in man with the use of a flow directed balloon-tipped catheter. N Engl J Med 1970;283:447–451.
22. Ganz W, et al. A new technique for measurement of cardiac output by thermodilution in man. Am J Cardiol 1971;27:392–396.
23. Cole J, Martin WE, Cheung PW, Johnson CC. Clinical studies with a solid state fiberoptic oximeter. Am J Cardiol 1972;29: 383–388.
24. Morris AH, Chapman RH, Gardner RM. Frequency of wedge pressure errors in the ICU. Crit Care Med 1985;13:705–708.
25. Cengiz M, Crapo RO, Gardner RM. The effect of ventilation on the accuracy of pulmonary artery and wedge pressure measurements. Crit Care Med 1983;11:502–507.
26. Rizvi K, et al. Effect of airway pressure display on interobserver agreement in the assessment of vascular pressure in patients with acute lung injury and acute respiratory distress syndrome. Crit Care Med 2005;33:98–103.
27. Liebowitz AB. More reliable determination of central venous and pulmonary artery occlusion pressures: Does it matter? Editorial Crit Care Med 2005;33:243–244.
28. Stevens JH, et al. Thermodilution cardiac output measurements: Effects of the respiratory cycle on its reproducibility. J Am Med Assoc 1985;253:2240–2242.
29. Pinsky MR, Vincent J-L. Lets use the pulmonary artery catheter correctly and only when we need it. Point of view. Crit Care Med 2005;33:1119–1122.
30. Levin PD, Sprung SL. Another point of view: No Swan song for the pulmonary artery catheter. Crit Care Med 33:1123–1124.

31. Gibbs NC, Gardner RM. Dynamics of invasive pressure monitoring systems: Clinical and laboratory evaluation. Heart Lung 1988;17:43–51.
32. Gardner RM, Hujcs M. Fundamentals of physiologic monitoring. AACN Clinical Issues Crit Care Nursing 1993;4(1):11–24.
33. Daily EK. Hemodynamic waveform analysis. J Cardiovasc Nurs 2001;15(2):6–22.
34. McGhee H, Bridges EJ. Monitoring arterial blood pressure: What you may not know. Crit Care Nurse 2002;22:60–78.
35. Sander O, et al. Impact of prolonged elevated heart rate on incidence of major cardiac events in critically ill patients with a high risk of cardiac complications. Crit Care Med 2003; 33:81–88.
36. Weissman C, Landesberg G. Vital are the vital signs. Comment/Editorial Crit Care Med 2003;33:241–242.
37. Grogan EL, et al. Volatility: A new vital sign identified using a novel bedside monitoring strategy. J Trauma 2005;58:7–14.
38. Poon KB. Fusing Multiple Heart Rate Signals to Reduce Alarms in Adult the Intensive Care Unit. MS dissatation, University of Utah Department of Medical Informatics, Salt Lake City (UT); May 2005.
39. Gardner RM. Information management hemodynamic monitoring. Semin Anesth 1983;2:287–299.
40. Gardner RM. Computerized management of intensive care patients. MD Comput 1986;3(1):36–51.
41. Shabot MM, Carlton PD, Sadoff S, Nolan-Avila L. Graphical reports and displays for complex ICU data: A new, flexible and configurable method. Comput Methods Programs Biomed 1986;22:111–116.
42. Gardner RM, Huff SM. Computers in the ICU: Why? What? So what? Intl J Clin Monit Comput 1992;9:199–205.
43. Gardner RM, Sittig DF, Clemmer TP. Computer in the ICU: A match meant to be! In: Ayers SM, Grenvik A, Holbrook PR, Shoemaker WC. editors. Textbook of Critical Care Medicine. 3rd ed. Philadelphia: W. B. Saunders;1995: p 1757–1770. chapt. 196.
44. Morris AH. Treatment algorithms and protocolized care. Curr Opin Crit Care 2003;9:236–240.
45. Morris AH. Clinical trial of a weaning protocol. Crit Care 2004;8:207-209.
46. Bria II WF, Shabot MM. The electronic medical record, safety and critical care. Crit Care Clin 2005;21:55–79.

Further Reading

Nichols WW, O'Rourke MF. McDonald's Blood Flow in Arteries: Theoretical, Experimental and Clinical Principles. 5th ed. Of special note is Chapter 6: Measuring principles of arterial waves pages 132–135. Hodder Arnold; 2005.

Daily EK, Tilkian AG. Hemodynamic Monitoring. In: Tilkian AG, Daily EK, editors. Cardiovascular Procedures: Diagnostic Techniques and Therapeutic Procedures. St. Louis (MO): Mosby; 1986 chapt. 4.

Geddes LA. The Direct and Indirect Measurement of Blood Pressure. Chicago: Year Book Medical Publisher; 1970.

Geddes LA. Handbook of Blood Pressure Measurement. Clifton (NJ): Humana Press; 1991.

Webster JG, editor. Design of Pulse Oximeters. Institute of Physics; 1997.

Webster JG, editor. Medical Instrumentation: Application and Design. 3rd ed. New York: John Wiley & Sons Inc.; 1998.

Tungjitkusolmun S, Heart and Circulation, In: Webster JG, editor. Bioinstrumentation. John Wiley & Sons, Inc.; 2004. Chapt. 8.

Shabot MM, Gardner RM, editors. Decision Support Systems for Critical Care. New York: Springer-Verlag Inc.; 1994.

Gardner RM, Shabot MM. Patient-monitoring systems. In: Shortliffe EH, Cimino JJ, editors. Biomedical Informatics: Computer Applications in Health Care. 3rd ed. New York: Springer-Verlag; 2005; in press, Aug. 2005. Chapt. 17.

See also Blood pressure measurement; cardiac output, fick technique for; cardiac output, indicator dilution measurement; cardiac output, thermodilution measurement; electrocardiography, computers in; hemodynamics.

MONITORING, INTRACRANIAL PRESSURE

Michael L. Daley
Ian Piper
The University of Memphis
Memphis, Tennessee

INTRODUCTION

Intracranial pressure (ICP) monitoring is a form of pressure monitoring used in patients with pathologies that might give rise to raised ICP. The definition of raised ICP depends on the underlying pathology and, for example in adult patients who have sustained a severe head, injury is defined as pressure greater than 20 mm Hg. Over the past 50 years there has been an active and wide ranging research into the causes and consequences of raised ICP that, to date, has been the subject of 11 international symposia embracing such diverse disciplines as neurosurgery, anesthesia, radiology, biophysics, electrical and mechanical engineering, mathematics, and computer science. This article reviews the underlying physiology pertinent to measurement of ICP as well as gives an overview of some of the current technology used to measure ICP. Then, a brief review follows of the clinical literature underlying the case for ICP measurement, with an emphasis on the main clinical condition of patients with a head injury. The last two sections focus on the use of waveform analysis and mathematical modeling techniques to research into mechanisms underlying raised ICP.

PHYSIOLOGY

ICP is the pressure recorded within the tissue (parenchyma) or fluid-filled spaces and is not uniformly distributed within the craniospinal axis. The craniospinal axis consists of all neural tissue and contiguous fluid-filled spaces within the cranium and spinal sac. The central neural tissue is encapsulated within bone and three-layered tissue coverings or *meninges*. From outside in the meninges are the dura, arachnoid, and pia membranes. The meninges provide a physical barrier between neural tissue and the external environment but also serve both a structural and a physiological function as an architecture for supporting cerebral vessels and maintaining a space for flow of cerebrospinal fluid (CSF). CSF cushions the delicate neural tissues, supports the weight of the brain, and acts as a transport media for nutrients, chemical messengers, and waste products. Except at the choroid plexus, the CSF is freely permeable to the appendymal lining and is in chemical communication with the interstitial fluid and the CNS. ICP is a reflection of the relationship between alterations in craniospinal volume and the ability of the craniospinal axis to accommodate added volume. The craniospinal axis is essentially a partially closed box with container properties including both viscous and elastic elements. The elastic or, its inverse, the compliant properties of the container will determine what added volume can be absorbed before ICP begins to rise. So an understanding of raised ICP encompasses an analysis of both intracranial volume and craniospinal compliance.

The history of the subject of ICP has been well reviewed (1,2) and starts from the doctrine named after Monro (3) and Kellie (4), which proposed that the brain and its contained blood were incompressible, enclosed in a rigid and inextensible skull, of which the total volume remained constant. In its original form, the Monro–Kellie doctrine did not take into account the CSF as a component of the cranial compartment. The concept of reciprocal volume changes between blood and CSF was introduced in 1846 by Burrows and later extended in the early twentieth century by Weed et al. (5,6) to allow for reciprocal changes in all craniospinal constituents. Kocher (7) in 1901 translated into clinical terms the four stages of cerebral compression proposed almost 25 years earlier by the experimental studies of Duret (8). Kocher described four stages of cerebral compression related to the expansion of intracranial brain tumours. In stage 1, the initial increase in tumor volume is compensated by a reduction in volume of the other intracranial components, chiefly CSF and venous blood. This spatial compensation results in no net increase in intracranial volume or pressure and hence no clinical symptoms. In stage 2, the compensatory mechanisms are exhausted, ICP increases, and the patient becomes drowsy with a headache. Stage 3 is characterized by a considerable increase in ICP, an associated deterioration in conscious level, and intermittent elevations of blood pressure (BP) accompanied by bradycardia. In the fourth and final stage, the patient is unconscious, with bilateral fixed dilated pupils and falling BP usually leading to death.

Cushing (9–11), then a research worker for Kocher, described in both experimental and clinical studies the close relationship between increases in ICP and BP and proposed that the BP rose to maintain adequate blood supply to the hind brain, the stimulus to this vasopressor response believed to be medullary ischemia (12,13). At about this time, a false confidence developed in the lumbar CSF pressure technique (lumbar puncture), which caused Cushing's findings to be challenged. Reports emerged (14–16) that some patients showing clinical signs of brain compression had normal lumbar CSF pressures and that in other patients elevations in BP were found at times when ICP was well below the level of BP.

Partly because of this apparent dissociation between ICP and clinical symptoms, emphasis switched away from ICP measurement toward the relationship between craniospinal volume and pressure, particularly the importance of the elastic properties of the craniospinal system. The relationship between ICP and intracranial volume is often described graphically by an exponential function that is relatively flat at low ICPs but rises rapidly as pressure increases much above 20 mm Hg. Under normal conditions, at the low ICP end of this curve, compliance (the ratio of added volume to pressure) is high. When the patient's volume–pressure status changes to move further up the curve, compliance will fall rapidly. Measurement of craniospinal compliance in brain-injured patients may, therefore, offer the potential for early detection of raised ICP before it rises to levels damaging to brain parenchyma. The most commonly used methods of measuring craniospinal compliance were developed by Marmarou (17,18) and depend on the rapid injection of known volumes of fluid into the CSF space with immediate measurement of the resultant increase in CSF pressure. One of these methods is the pressure volume index (PVI); this value being the volume that when added to the CSF space would produce a tenfold rise in ICP. Miller et al. (19,20) defined a further measure of the craniospinal volume–pressure relationship, the volume pressure response (VPR). The VPR, also calculated from the ICP response resulting from a rapid bolus injection of saline into the CSF space, is a direct measure of the inverse of compliance: elastance. Although both are measures of compliance, some confusion remains as to what exactly is the difference between the VPR and the PVI. The PVI, which assumes a mono-exponential pressure versus volume relationship, is a single index that characterizes the patient's entire volume–pressure relationship and is, numerically, a measure of the slope of the logarithm of the ICP versus intracranial volume relationship. Its strength lies in its ability to define the whole volume–pressure status of the patient with a single index. It was the late J. Douglas Miller who pointed out that if there was only a single volume–pressure curve, then no new information would be gained by measuring compliance or elastance, and a knowledge of absolute ICP alone would suffice in determining the state of a patient's craniospinal volume decompensation. However, several studies (21,22) have shown that the shape of the volume–pressure relationship changes under a variety of conditions both between patients and within patients at different times. Under these circumstances, it is likely that the PVI will provide the more useful information. One weakness of the PVI is that if a patient's pressure–volume relationship remains stable, single measurements will not detect movement along a given pressure–volume curve and thus may not detect slow increases in volume of a space-occupying lesion before it causes significant increases in ICP. It is in this situation where a continuous measure of absolute compliance (or its inverse: elastance as measured by the VPR) provides, potentially, more useful information. It is for this reason that, in head-injured patients, Miller et al. found that the VPR correlated better to the degree of brain mid-line shift, as imaged on CT scan, than it did to absolute ICP alone. They subsequently demonstrated that the VPR could serve as an indicator for surgical decompression, critical levels being between 3 and 5 mm Hg/mL (19,23). However, in clinical practice, both circumstances (movement along the pressure–volume curve and shift of the curve along the pressure axis) may occur within a patient at different times and thus shows the need for ICP monitoring capable of both continuous compliance measurement and derivation of the PVI.

Few clinicians would argue over the theoretical utility of monitoring compliance in brain-injured patients; what has limited its use in practice has been the inherent problems associated with manual volume–pressure testing. It is difficult to inject equal volumes of fluid manually at a constant and rapid rate of injection. As a result, repeated measures are usually required that can result in a lengthy and time-consuming procedure. Also, even with stringent maintenance of the sterile procedure, repeated addition or removal of CSF from patients carries an increased risk of infection to the patient. Despite the reported benefits of compliance measurement, it has been largely these limitations that have prevented widespread adoption of compliance measurement in clinical practice.

It was not until the 1960s when Lundberg (24) published his now classic monograph that interest in clinical ICP measurement was rekindled. Using ventricular fluid pressure recording in brain tumor patients, Lundberg was the first to delineate the frequency with which raised ICP occurs clinically, at times reaching pressures as high as 100 mm Hg. Lundberg also described three types of spontaneous pressure wave fluctuations: "A" waves or plateau waves of large amplitude (50–100 mm Hg) with a variable duration (5–20 min), "B" waves that are smaller (up to 50 mm Hg), and sharper waves with a dominant frequency of 0.5–2 cycles per min.

DEVICES

ICP was first measured experimentally in animals by Baylis (25) in 1897 using an early form of strain gauge. The most common form of strain gauge use specially prepared alloys that change their resistance in proportion to the amount they are elongated or stretched in cross section due to applied strain.

The Wheatstone Bridge and the Strain Gauge

Most common strain gauge transducers are used in a Wheatstone Bridge configuration, where the resistive strain elements are placed on diagonally opposite arms of the bridge. Should the strain gauge change its resistance (for example, due to an applied strain—perhaps caused by a pressure acting on the strain gauge to cause it to elongate), then the bridge will become unbalanced and a potential difference will be generated in proportion to the degree of strain.

Fluid-Filled Catheter Transducer Systems

The standard intraventricular catheter connected to an external strain gauge transducer is called a catheter-transducer system because it behaves, in many ways, like a mechanical system with a *mass* of fluid that acts against the spring-like "elastic" properties of the catheter walls and the transducer diaphragm. A typical catheter-transducer system is one used for the measurement of arterial pressure comprising a silicon strain gauge, fluid-filled catheter, three-way tap, and an arterial cannulae. Modern transducers are semiconductor fabricated and normally disposable. If a flushing device is attached to the transducer, extreme care must be used to avoid accidental flushing into the cranium. The older catheter-transducer system included long tubing and therefore had different frequency characteristics than modern transducers. The older transducer first found widespread use in the 1960s and 1970s following the pioneering work on long-term ICP monitoring by Nils Lundberg (24). Most publications in this field still refer to intraventricular monitoring as the "Gold-Standard" method of measuring ICP for several reasons. First, this method allows checking for zero and sensitivity drift of the measurement system *in vivo*. Second, pressure measurement within the CSF space transduces pressure within a medium that is an incompressible fluid and, provided CSF flow is not blocked, is not subject to the development of intracompartmental pressure gradients. Finally, access to the CSF space provides a method for ICP treatment via CSF drainage. However, concerns are often expressed about the increased risks of infection associated with ventriculostomy. Although a range of infection rates has been reported, some as high as 40% (26), recent reports confirm infection rates to be in the region of 1%, which is not considered a prohibitive risk (27).

The Head Injury Management Guidelines published by the Brain Trauma Foundation in 1994 recommend intraventricular ICP measurement as the first-line approach to monitoring ICP. They state that "A ventricular catheter connected to an external strain gauge transducer or catheter tip pressure transducer device is the most accurate and reliable method of monitoring ICP and enables therapeutic CSF drainage. Clinically significant infections or haemorrhage associated with ICP devices causing patient morbidity are rare and should not deter the decision to monitor ICP" (28). Despite the existence of these guidelines (27,28), catheter-tip intraparenchymal pressure monitoring remains popular, particularly in the United Kingdom, as it does not require catheter placement in the operating theater and thus carries significantly lower resource implications. However, the routine use of fluid-filled catheter-transducer systems is not without difficulties. A catheter-transducer system can be described as a second-order mechanical system and, if under-damped, will oscillate at its own natural frequency producing significant amplitude and phase distortion of the pressure signal. The degree of distortion will depend on the damping factor (β) of the system. For most purposes, a damping factor of 0.64 is optimal as the amplitude error will be less than 2% for up to two thirds of the natural frequency of the system (29). Typically, most pressure catheter-transducer systems used in patients tend to be under-damped with a damping factor (β) less than 0.4 (29,30). Manipulating a catheter-transducer system from under-damped ($\beta < 0.3$) to over-damped ($\beta > 0.8$) can cause a decrease in mean pressure of 7 mm Hg. Commercial devices are available, however, such as the acudynamic adjustable damping device (31) that can alter the damping characteristics of external strain gauge pressure transducers and bring them within the range of optimal damping. Another problem with fluid-filled catheter-transducer systems is correcting for the presence of hydrostatic pressure gradients when measuring cerebral perfusion pressure (CPP). Typically, the external strain gauge transducer is zeroed at the same level as the arterial pressure (BP) transducer, usually at the level of the right atrium. Although the patient is managed in the horizontal position, there is no column of fluid between the site of ICP and BP measurement. However, when the patient is managed with head-up tilt and if the BP transducer is not moved to the same horizontal level as the head, a hydrostatic fluid column will be created. This can produce a significant error between the observed CPP and the actual CPP, which, in the worst case, can produce an error of as much as 15 mm Hg.

The Spiegelberg ICP monitoring system (Spiegelberg KG, Homburg, Germany) largely overcomes these problems. This system is a special case of a fluid-filled catheter-transducer system. With this device, ICP is measured using a catheter with an air pouch balloon situated at the tip. By maintaining a constant known volume within the air pouch, the pressure within the air pouch balloon is equivalent to the surrounding pressure or ICP. The internal air pouch balloon is transduced by an external strain gauge transducer, and because the fluid used for pressure transduction is air, the pressure error caused by an "air column" is clinically insignificant. The design of this device also allows automatic *in vivo* zeroing of the ICP system and, in laboratory bench tests, showed the least zero drift in comparison with standard catheter-tip ICP devices (32). This system now has versions for use in epidural, subdural, intraparenchymal, and intraventricular sites. The intraventricular catheter is a double lumen catheter that allows access to the CSF space for drainage. The Spiegelberg system, although having many attractive features, is, as yet, a relatively little used system outside of Germany and its long-term clinical utility and robustness require evaluation.

Catheter-Tip Transducer Systems

Now several catheter-tip ICP monitoring systems are available, including the Camino (33–36) systems. The Gaeltec ICP/B solid-state miniature ICP transducers, for use in the epidural space, are reusable, and the zero reference can be checked *in vivo*. However, reports of measurement artifacts (37) and decay in measurement quality associated with repeated use (35) have limited the widespread adoption of this technology.

The InnerSpace OPX 100 system (InnerSpace Medical, Irvine, CA) is, like the Camino system, a fiber optic system. Bench test reports on this system show it to have good zero drift and sensitivity stability (32). However, a recent clinical evaluation of this system in 51 patients reported a high (17%) incidence of hematoma formation around the ICP sensor (36). The authors concluded that improved fixation of the catheter is required to minimize micro-movements.

The two catheter-tip systems most frequently used in the management of head-injured patients are the Codman and Camino systems. Neither allows a pressure calibration to be performed *in vivo*. After these systems are "zeroed," relative to atmospheric pressure during a pre-insertion calibration, their pressure output is dependent on zero drift of the sensor. For this reason, it is critical that these devices exhibit good long-term zero drift characteristics. These devices provide an electrical calibration, to calibrate external monitors, but they cannot be corrected for inherent zero drift of the catheter once placed. For the Camino system, the manufacturers specify the zero drift of the catheter to be ±2 mm Hg for the first day and ±1 mm Hg per day thereafter. Czosnyka et al. (32) have confirmed these zero drift findings in bench tests studies although they also reported that the temperature drift of the device was significant (0.3 mm Hg/°C). They reported that if the manufacturers specify the zero drift of the catheter to be ±2 mm Hg for the first day and ±1 mm Hg per day thereafter. In clinical practice, the reported zero drift upon removal of the Camino device from the patient has been reported to be greater than the manufacturer's specifications. Munch (38) assessed 136 Camino sensors in a clinical study and found an average *daily* drift rate of 3.2 mm Hg. Chambers (39), in a comparative study of the Camino ventricular catheter with an external fluid-filled catheter-transducer system, reported that only 60% of the readings were within 2 mm Hg of the gold-standard method. There are also reports of Camino probe failure because of technical complications (cable kinking, probe dislocation), with reported failure rates ranging from 10% to 25% (39,40).

The Codman transducer is a micro-miniature strain gauge within a titanium housing side mounted at the tip of a catheter. Similar to other transducers, bench test reports on this technology have been favorable (32,41). However, clinical evaluations have reported the presence of inter-patient and intra-patient biases that are independent on whether the device is compared against the Camino transducer or an intraventricular catheter-transducer system (42). A report by Fernades (43) found that in 24% of the recordings, the Codman sensor over-read the Camino system by 5 mm Hg or more.

The Rehau System

The technology of miniaturization is allowing the development of catheter-tip pressure sensors. Catheter-tip systems, due to their small diameter, are likely to cause less damage to tissue upon placement than larger fluid-filled catheters and are not affected by hydrostatic pressure differences. However, they are potentially more prone to problems of robustness and *in vivo* zero drift. Several similar technologies, although performing well in bench test studies, have been shown to exhibit unacceptable zero drift, fragility, or both during trials conducted under clinical conditions (38,43–45). *In vivo* drift is especially important in catheter-tip strain-gauge technology as it is impossible to check if the calibration has altered after being placed in the patient. Until recently, to reduce the physical size of the catheter-tip strain gauge catheters, often only a partial Wheatstone bridge is employed at the catheter-tip. Recently, a new technology has become available, the Neurovent-P (REHAU AG+CO, REHAU, Germany), in which a full Wheatstone bridge is fabricated in the probe tip, and this solution should, in theory, provide improved zero drift characteristics. One potential advantage of the Rehau NeuroVent system seems to be the incorporation of a full Wheatstone bridge into the catheter-tip electronics. This Wheatstone technological improvement should enhance the zero drift characteristics by reducing temperature sensitivity and the effects of non-pressure-related external strains. Until recently, this approach was only possible in the much physically larger external strain gauge systems. Through advances in miniaturization technology, it has now been able to incorporate this technology into its catheter-tip systems. In the bench studies performed (46), each catheter was tested for days, from a minimum of 3 to more than 8 days. This timeframe typically covers the period in which an ICP transducer is used in the clinical setting. The results from this bench test study confirm the manufacture's claim that pre-insertion calibration is not required as the drift was within manufacture's specifications (±1 mm Hg). This advantage is very important in the clinical arena because the first pre-insertion zeroing is a crucial step conducted by the surgeons and potentially, if incorrectly performed, could generate erroneous readings during the period of monitoring. Both long-term zero drift and the dynamic pressure test results also confirm that this system performs well in bench test studies and meets manufacturer's specifications. Mean zero drift, after 5 days, was very small and long-term continuous recording of a stable pressure was precise. The response to dynamic tests, i.e., the changes of pressure over a wide range, was excellent. The average bias of the Rehau catheter compared with a hydrostatic pressure column is very small. Despite these promising bench test study results, further work is required to determine the performance of this measurement device in the clinical environment. Following on from these bench tests, the next and most critical step will be to conduct a trial of this promising technology under the more demanding clinical environment. The BrainIT group (http://www.brainit.org) as a multicenter collaborative group of neurointensive care scientists and clinicians are well placed to design and conduct such a trial (47). Should this technology demonstrate, under clinical conditions, the required robustness and low drift as indicated by these bench test studies, it may lead to more precise and reliable measurement of ICP.

CLINICAL LITERATURE

Raised ICP has been found to be associated with a poorer outcome from injury with the higher the level of ICP, particularly the peak ICP level, which has been found to correlate with the expected prognosis for mortality and morbidity (48–51). There has, however, been controversy over the usefulness of monitoring raised ICP with some groups, with a "no ICP monitoring" policy, finding in their studies of head injury mortality and morbidity that outcome is similar to other groups that do monitor ICP (52). Reported differences in the utility of ICP monitoring could be due to variability both in management and in monitoring protocols between different neurosurgical centers. Variation in type of ICP pressure monitor, site of placement, treatment thresholds, patient referral characteristics, and in outcome measures can all combine to produce a large variability both in measured ICP and in outcome irrespective of whether ICP is monitored or how it is treated. Another source of variation in terms of raised ICP is the inherent variability of the head-injured population with outcome being dependant on several other factors. For example, mass lesions are generally accompanied by elevations in ICP of greater than 40 mm Hg and are associated with poorer outcome, whereas diffuse injuries tend to have lower ICP levels associated with a similar poor outcome (50,53). Age is also an important factor with an age-dependent distribution of ICP for both type of injury and outcome. This is particularly so for pediatric cases (54–56). ICP can even be raised in the absence of overt signs of swelling or mass lesions on CT. In a small study of severely head-injured patients, O'Sullivan (57) demonstrated that some comatose head-injured patients whose initial CT scan was normal, with no mass lesion, midline shift, or abnormal basal cisterns, developed raised ICP greater than 20 mm Hg that lasted longer than 5 min. This included a subset of patients showing pronounced raised ICP of greater than 30 mm Hg.

Data from large prospective trials carried out from single centers and from well-controlled multicenter studies have provided the most convincing evidence for a direct relationship between ICP and outcome (58–61). Narayan et al. (58) in a prospective study in 133 severely head-injured patients demonstrated that the outcome prediction rate was increased when the standard clinical data such as age, Glasgow Coma Score on admission (GCS), and pupillary response with extraocular and motor activity was combined with ICP monitoring data. Marmarou et al. (60), reporting on 428 patients' data from the National Institute of Health's Traumatic Coma Data Bank, showed that following the usual clinical signs of age, admission motor score, and abnormal pupils, the proportion of hourly ICP recordings greater than 20 mm Hg was the next most significant predictor of outcome. Outcome was classified by the Glasgow Outcome Score (GOS) at 6 months follow-up. They also found, using step-wise logistic regression, that after ICP, arterial pressure below 80 mm Hg was also a significant predictor of outcome. Jones et al. (61) studied prospectively 124 adult head-injured patients during intensive care using a computerized data collection system capable of minute by minute monitoring of up to 14 clinically indicated physiological variables. They found that ICP, above 30 mm Hg, arterial pressure below 90 mm Hg, and cerebral perfusion pressure below 50 mm Hg significantly affected patient morbidity.

Although differing opinions remain about the contribution of continuous monitoring of ICP to reduction in mortality and morbidity after head injury, there is now sufficient evidence to remove doubt about the value of ICP monitoring toward improving the detection and preventative management of secondary cerebral injury.

Raised Intracranial Pressure: Relationship to Primary and Secondary Injury

Both experimental and clinical studies have clearly shown that after traumatic brain injury, normal physiological mechanisms for maintaining cerebral perfusion can become impaired (62–65). These studies demonstrate that brain injury can cause impairment or loss of autoregulation defined as the ability of the cerebral vessels to respond to changes, in arterial gases or to arterial pressure. As a result of these changes, there can, at times, be a decrease in cerebrovascular resistance that can lead to raised ICP in both adults and children (66–70). Although brain-injured patients are being managed in intensive care, there are, superimposed on to the primary injury, periods of reduced arterial PO2 or episodes of arterial hypotension often as a result of other injuries or treatment by hypnotic drugs (50,53,60,61,71,72). With an impaired physiological mechanism unable to respond adequately to these adverse changes in physiological parameters (or "secondary insults"), ischemic brain damage can occur. These secondary, chiefly ischemic brain insults are common, with Graham et al. (74) reporting, in a series of 151 fatal cases of severe head injury, a 91% incidence of ischemic brain damage found on autopsy. A second study carried out by the same group over 10 years later found a similar high incidence (>80%) of ischemic brain damage despite subsequent improvements in intensive care of head-injured patients (74).

There has been much interest in the relationship among ICP, CPP, and CBF. The landmark study of Miller and Garibi (75) produced some of the first experimental evidence confirming the concept that changes in ICP affect cerebral blood flow not directly but through changes in CPP, where CPP is defined as the difference between mean arterial pressure and ICP. Strictly speaking, the actual cerebral perfusion outflow pressure would be cerebral venous pressure, although this pressure is, in most situations, impractical to measure routinely. However, it has been established that over a wide range of pressures cerebral venous pressure is well approximated (within 3–4 mm Hg) by ICP (76–78). In an experimental study of CBF as determined by the venous outflow technique in dogs, Miller and Garibi (75) also demonstrated that when MAP and ICP rise in parallel so that CPP remains constant at 60 mm Hg, CBF increases with MAP in animals found to be non-autoregulating. It was further shown that as CPP drops in autoregulating animals, the breakpoint at which CBF starts to decrease is at a higher level if CPP is reduced through hemorrhagic arterial hypotension than through intracranial hypertension. This work suggests that cerebral perfusion is more sensitive to arterial hypotension than to intracranial hypertension.

The clinical significance of this information is that in the management of head injury, it is often necessary to employ therapy to lower raised ICP. Therapeutic agents for reducing raised ICP often do so at the expense of reduced MAP, and as a consequence, CPP may not improve. If autoregulation is preserved, CBF should remain unchanged despite parallel changes in MAP and ICP. However, clinically, autoregulation is likely to be impaired in those conditions in which ICP is increased such as head injury or subarachnoid hemorrhage (68,69,79–82). Under these circumstances, it is important that reduction in ICP should not be achieved at the expense of lowering CBF and provoking brain ischemia.

This earlier work of Miller and Garibi was later extended by Chan et al. (83) to include CPP ranges of 60, 50, and 40 mm Hg. At CPP levels of 50 and 60 mm Hg, when autoregulation was intact, CBF remained unchanged. However, with loss of autoregulation, there was a trend for CBF to increase as MAP and ICP were increased in parallel at a CPP of 50 and 60 mm Hg. Absolute CBF levels were significantly different between the autoregulating and non-autoregulating groups. At a CPP of 40 mm Hg CBF showed a linear correlation with BP. This work demonstrates that when autoregulation is impaired, there is a functional difference between autoregulating and non-autoregulating cerebral vessels despite similar MAP and CPP, and that when autoregulation is impaired, CBF depends more on arterial driving pressure than on cerebral perfusion pressure.

The importance of arterial pressure as the prime factor governing CPP-related secondary insults has been well demonstrated by the work of Jones et al. (61), where they carried out a prospective study over 4 years of the frequency and severity with which secondary insults occur to head-injured patients while being managed in intensive care. They developed a microcomputer-based data collection system that allows the acquisition of data from up to 15 monitored variables minute by minute (84). At each bed space, data collection was under the control of a microcomputer where serial links between the patient monitors and the microcomputer allow the controlled transfer of multiple channels of physiological data once per minute. The controlling software allowed medical staff to add comments to the current active computer file at any time, precisely annotating significant events. The software performs artifact detection, calculates derived data, and highlights valid data that falls outside normal physiological levels. Collected data were stored to disk and could be printed either locally or remotely. Later work by Chambers et al. (85) found that both raised ICP and lowered ABP are major factors in producing secondary insults.

From the valid physiological data produced, manual processing of data was used to identify secondary insults that are defined at one of three grades of severity and that must last for 5 min or longer to be recorded as an insult. This permits calculation of the frequency, severity, and total duration of insults, measured in minutes.

An analysis was made of 124 adult head-injured patients who were monitored during intensive care using the computerized data collection system. Information was logged at 1 min intervals and scanned to identify insults when values fell outside threshold limits for 5 min or longer. Three grades of insult were defined for each variable (Table 1). The duration of insults has been analyzed in relation to the GOS of these patients at 12 months after injury. The monitored patients included 68 with severe head injury (GCS 8 or less with no eye opening), 36 with moderate head injury (GCS 9–12), and 20 with minor head injury (GCS 13–15 but with multiple injuries, scoring 16 or more on the Injury Severity Scale).

Insults were found in 91% of patients at all degrees of severity of head injury. Overall, 10% of patients had insults that were only at the lowest grade 1 level, 31% had insults at both grade 1 and grade 2 levels, and 50% of patients had at least one insult at grade 3 level in addition to grades 1 and 2 insults. Overall, the majority (77%) of all insults detected in the ITU were at grade 1 level, and these represented 85% of the total duration in minutes of insult measured. Differences in the duration of insult between outcome groups 12 months post-injury were compared with each grade of insult using Kruskall–Wallis one-way analysis of variance and Mann–Whitney U tests. Significant differences in the distribution of hypotensive insults were found between the outcome grades at all levels of severity of insult. Similar results were found for cerebral perfusion pressure insult duration. These data confirm the important adverse effect of even moderate reductions in arterial pressure (systolic BP less than 90 mm Hg or mean BP less than 70 mm Hg).

Although the occurrence and clinical significance of severe and long-lasting secondary insults in head-injured patients is not disputed. The incidence, severity, and duration of shorter acting "minute by minute" cerebral perfusion pressure insults, as defined by the Edinburgh secondary insult detection methodology, has not been defined outside of the Edinburgh study population. In addition, the strong association between the occurrence of specific insult types and the subsequent patient morbidity and mortality found by the Edinburgh study (61) needs to be reproduced in other centers. From the Edinburgh group, Signorini et al. (86,87) developed and validated a model for predicting survival in head-injured patients based on collection of simple demographic features. When the minute by minute secondary insult data were added to the baseline model, they found only ICP insults significantly improved the fit of the model.

Almost all of the evidence for a CPP management is based on single-center cohort studies, often compared with historical controls. For example, Rosner and Becker (26) reported on clinical results of a CPP management protocol where approximately 40% of patients received vasopressor support. They reported a greatly improved incidence of favorable outcome in patients' $GCS \geq 7$ compared with historical controls. Despite evidence from single-center studies as described above, a critique of the literature for the purposes of defining head injury management guidelines published by the Brain Trauma Foundation in 1994 states that there is not sufficient evidence to establish either a standard or a guideline for the management of CPP; however they indicate management of CPP greater than 70 mm Hg as a management option (28). Since then, Robertson et al. (88) performed one of the first randomized controlled single-center trials of the CPP management approach. They defined two management cohorts, one based on their normal practice of CPP management and the other CBF, guided practice that included the aggressive management of CPP above 70 mm Hg, together with restricted use of significant hyperventilation. Conversely, their trial has shown that aggressive management of $CPP > 70$ mm Hg, although reducing the incidence of jugular venous desaturation $< 50\%$, demonstrated no difference in neurological outcome possibly due to the increased incidence of acute respiratory distress syndrome in the CBF management group. Thus, secondary insults are common, result in mainly ischemic brain damage, and are a major contribution to disablement. Moreover Chambers et al. (89) has reported that the critical CPP thresholds for insults of pediatric patients varies with age. For both adult and pediatric patients, insults are important because they are common and yet so potentially avoidable. Clearly a critical challenge facing us is to develop patient monitoring systems and protocols that will lead to rapid detection and resolution of secondary insults. However, detection is not enough; we need also improved and clinically proven methods of treating secondary insults—further evidence is required. In Signorini et al.'s article (87), they conclude that the questions posed by such observational studies can only be answered definitively within the context of a randomized clinical trial. However, to design such a multicenter randomized clinical trial will require improved standards in the monitoring and analysis of secondary insult data. In Europe, improved standards for high-resolution collection and analysis of multicenter data from head-injured data is now being addressed by the *Brain-IT* group (90). The *Brain-IT* group (http://www.brainit.org) is an open consortium of clinicians and basic scientists working toward improving the infrastructure for conducting both observational and controlled trials of medical devices and patient management.

INTRACRANIAL PRESSURE ANALYSIS METHODS

Since the early 1990s, clinical monitoring of ICP has generally become part of the intensive care management of patients with brain injury. Several methods of analysis have been designed to extract pathophysiological information from the ICP and the corresponding ABP recordings. As noted, one particular computation used in clinical practice is the determination of mean CPP, the pressure across the brain. CPP is calculated as the difference between mean ABP and mean ICP and is based on the assumption that cerebral venous pressure is approximately equal to ICP. CPP is a useful parameter because it provides some insight as to whether the blood flow though brain capillaries is regulated. In the uninjured brain, cerebral blood flow is regulated to match the metabolic demand of the brain cells. Regulation of flow is primarily done by active dilation and constriction of cerebral arterioles in response to changes of CPP and/or biochemical vasoconstrictive or vasodilator agents that interact with the vascular endothelium. The steady-state relationship between cerebral blood flow and CPP is termed static pressure regulation and is illustrated by the graphical relationship between cerebral blood flow and CPP shown in Fig. 1.

In the steady state, cerebral blood flow is laminar and computed as the ratio of CPP to hemodynamic resistance, which is inversely proportional to the fourth power of the radius of the vessel. In the autoregulatory range, the arterial–arteriolar bed actively adjusts the resistance of its vessels by dilating when CPP decreases and constricting when CPP increases to maintain a relatively constant cerebral blood flow during changes in CPP. When CPP is below the lower limit of the autoregulatory range, vessels within the arterial–arteriolar bed tend to passively vasoconstrict. When CPP is above the upper limit of autoregulation, passive vasodilation occurs. One proposed intensive care therapeutic procedure designed to prevent secondary complications during recovery is CPP-oriented therapy (91). This therapy requires pressure autoregulation and the ability to manipulate CPP within the autoregulatory range (91). During intact pressure regulation, increases of

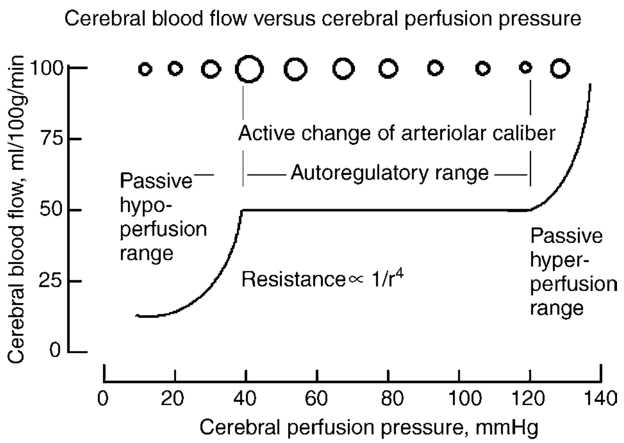


Figure 1. Illustration of static pressure regulation. Within the autoregulatory range, changes in arteriolar diameter, represented here as circles at the top of the graph, are primarily responsible for the regulation of cerebral blood flow. The arterial–arteriolar bed actively constricts with increasing CPP and dilates with decreasing CPP. Hypoperfusion occurs when CPP falls below the lower limit of regulation. As CPP falls below this limit, the vascompression of the arterioles and resistance increases markedly. The lower limit of regulation is not precisely known and varies with age, with 40–50 mm Hg for young pediatric cases (90) and 60–70 mm Hg for an adult (61,86).

CPP cause constriction of the arterial-arteriolar vascular bed and lowering of ICP by a reduction in cerebral blood volume. In addition, the resulting decrease of pre- and post-capillary pressure lessens fluid filtration and increases absorption, thus reducing the effects of edema. The application of CPP-oriented therapy when autoregulation has been lost may result in an imbalance of Starling forces at the capillaries, leading to increased net filtration and further brain injury by increased production of vasogenic edema.

In contrast to clinical cardiology, where the physiological mechanism underlying the dynamic features of the arterial pressure recording are fairly well understood, very little is known about the mechanism that underlies the shape of pulsations of intracranial pressure and how they are influenced by changes in CPP. Past studies have indicated that pulsations in the cerebrospinal fluid develop from either pulsations of the choroids plexus (92,93) or the arterial vasculature (94,95). Depending on the dilatory state of the cerebrovascular bed, ABP, ICP, and whether the autoregulation is intact, at times the primary component of the ICP pulse may be due to pulsations of venous or arterial origin (94–96). Although the origin of the ICP pulsation is not completely understood, possibly useful methodologies involving the pulsation have been developed. Over two decades ago, Avezatt and van Eijnhoven developed a procedure for distinguishing the occurrence of the loss of regulation of cerebral blood flow through numerical analysis of the pulsation of intracranial pressure (96,97) From laboratory studies, they found that within the autoregulatory range, the relationship between the mean amplitude of the pulsation of ICP increased linearly with mean ICP with a high slope value. At mean ICP above 30 mm Hg, the slope of this relationship decreased. They proposed that the flattening of the slope of this relationship was an indication of loss of autoregulation (97). A weakness of this technique is that it is dependent on heart rate (98). A low heart rate reflects an increase in the volume, of pulsatile cerebral blood inflow resulting from increased cardiac stroke volume, and the converse is likely for high heart rates. Thus, the variability in the amplitude of the ICP due to heart-rate variability can complicate the analysis (98). However, one modification of the analysis technique is to examine the correlation value between amplitude-pressure and mean ICP rather than the slope of the respective regression line (99). Positive values of this correlation value, which has been called RAP, indicate the ability of the vasculature to regulate cerebral blood flow, and negative values indicate poor cerebrovascular reserve and impaired cerebrovascular reactivity (99,100). Statistical analysis revealed that patients with a fatal outcome were associated with a negative RAP value (99). Most recently, the RAP index has been used to predict the achievable reduction in intracranial pressure a patient can obtain by the implementation of moderate hyperventilation (101). A modification of the mean amplitude of ICP and mean ICP characteristic has been proposed as a means of predicting which patients with idiopathic adult hydrocephalus will have a good outcome after shunting surgery (102). In this application, the amplitude of B-wave of ICP is used instead of the amplitude of the ICP pulsation (102).

In addition to relatively synchronous pulsations of ICP associated with the cardiac cycle, the ICP recording obtained during mechanical ventilation contains a low-frequency component at the rate of ventilation. Generally during mechanical ventilation, inhalation is produced by positive pressure and expiration occurs during zero pressure. Changes in pulmonary volume produce changes in intrathoracic pressure that mechanically modulate arterial and venous blood flow and produce a cyclic compression of the craniospinal sac. Evidence for this mechanical modulation by intrathoracic pressure can be observed in the spectra of an ICP recording obtained during mechanical ventilation. Because the pulsation in ICP associated with the cardiac cycle produced is quasi-periodic and the rate of ventilation is relatively constant, the spectra of the ICP pressure recording contains salient peaks at the cardiac frequency and its higher harmonics. In addition, each of the spectra associated with cardiac cycle has sidebands with a deviation at the ventilation frequency (103). Such a result is consistent with the premise that intrathoracic pressure changes mechanical modulated arterial and venous blood pressure and volume of the craniospinal sac (103).

Over the last few decades, several clinical and laboratory studies of the correlative relationship between ABP and ICP have been completed. Portnoy et al. reported that during normal vascular tone and intact regulation of cerebral blood flow, the ICP and ABP recordings do not look similar, but during maximum vasodilation of the arterial–arteriolar bed and impaired autoregulation induced by severe hypercapnia, these pressure recordings look remarkably similar (104–106). To numerically quantify the correlation between ABP and ICP, this group examined changes in the coherence function. Specifically, they found that the frequency domain coherence function approached unity when the ICP and ABP recording became similar. Generally, their observations and the observations of others using a spectral analysis systems approach have suggested that the more similar the spectral components of the ICP recording are to those of the ABP recording, the more likely cerebral autoregulation is impaired (106–108). The physical process describing the ABP as the input to the craniospinal sac and the ICP as the corresponding output has been termed cerebrovascular pressure transmission (106). An observational clinical study determined that four types of frequency descriptions of the transmission characteristic could be identified. Two types were associated with high ICP and the other two with low ICP (47).

More recent studies using time-domain correlation analysis on the ICP and ABP pressure recordings have been completed. As correlation analysis of two signals in the time domain is analytically equivalent to coherence analysis in the frequency domain, it was not unexpected that studies on the normocapnic/hypercapnic piglet model found that as the ICP and ABP recordings became more similar, the maximum value of the correlation function approaches unity, the pial arterioles became more dilated, and cerebral blood flow increased (109,110). Consistent with clinical reports that indicate that unlike adult patients, brain-injured pediatric patients often demonstrate cerebral hyperemia and increased ICP (82,109,111), correlation analysis of pressure recordings obtained from pediatric patients have been found to often approach unity, indicating the occurrence of inappropriate vasodilation and cerebral hyperemia (112). Most recently, Czosnyka et al. have reported the clinical use of a pressure reactivity index (PRx). This index is a moving correlation coefficient between 40 consecutive samples of values for intracranial and arterial pressures averaged for a period of 5 s (113,114). They have concluded that when slow waves in the ABP and ICP recordings are present, the proposed clinical index provides a continuous index of cerebrovascular reactivity (114,115). In particular, the PRx was designed to evaluate the integrity of the cerebrovascular response and estimate cerebrovascular autoregulatory reserve reactivity (114,115). The hypothesis is that because of the sluggish nature of the cerebrovascular system, naturally occurring slow-varying oscillations of ABP can be used to evaluate the autoregulatory reserve reactivity (114,115). Unlike the coherence and correlation indices described above, this index cannot be related to a linear system model. By employing averaging over a 5 s interval, most of the frequency changes above 0.2 Hz in the ABP and ICP recordings are filtered out. In addition, Nyquist's sampling theorem dictates that the highest frequency that can be represented by a signal sampled every 5 s is 0.1 Hz or six oscillations per minute. As a result, aliasing occurs, and the dynamical system relationship between ABP and ICP is not precisely defined by this index. Nevertheless, the PRx has been found to be a very useful tool. Clinical observations demonstrate that the PRx is high both during the occurrence of plateau waves and during refractory ICP hypertension (115). In addition, the PRx has been used to guide proposed therapies (116) and while variable seems to provide a reliable index of autoregulation (117).

Most recently, the regressive relationship between mean ICP and mean CPP has been used to assess whether autoregulation of cerebral blood flow is intact. In these studies, ABP is pharmalogically elevated and the regressive relationship between ICP and CPP during the test period is obtained (118). If pressure regulation of cerebral blood flow is intact, then increases of CPP will cause vasoconstriction, a decrease of ICP, and the regressive relationship should have a negative slope parameter. A marked positive slope parameter of this regressive relationship is an indication of passive pressure regulation; increases of CPP cause dilation and increased ICP (119).

ANALYSIS METHODS BASED ON MODELS OF ICP

Models of ICP dynamics are based on the modified Monroe–Kellie doctrine (see the Physiology section), which assumes that the total volume of intracranial substance, tissue, blood, and CSF, is constant. Initial mathematical models described the relationship between the formation of CSF and absorption of CSF in equilibrium. Specifically, in these models, laminar flow of CSF is assumed during steady-state conditions (120–123) and the parameters of the mathematical model are presented as an electrical circuit (Fig. 2).

Using this circuit and manipulating the volume of CSF by bolus either by withdrawal, injection, or constant infusion, it is possible to estimate R_o, C, the volume–pressure response, and the PVI. The latter two parameters have been discussed previously in the Physiology section. All parameters have been used to guide the management of hydrocephalous and traumatic brain injury.

Higher order mathematical models of intracranial hydrodynamics that incorporate the arterial and venous blood volume compartments into the analog electric circuit model have been developed to simulate the pulsatility, which is present in the ICP recording. To account for the reciprocal relationship between cerebral blood volume and volume of CSF, Agarwal et al. (124) constructed a model that connected vascular compliance to intracranial compliance in a series configuration. This modeling effort was further developed in detail by Ursino's proposed fourth-order analog circuit model of overall human intracranial

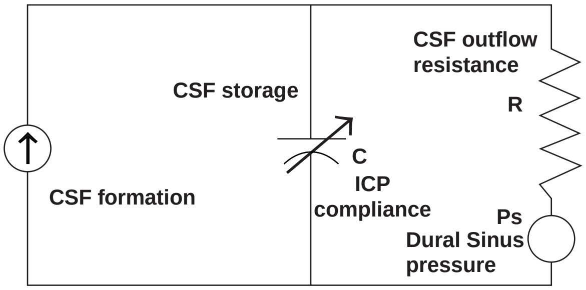


Figure 2. Electric circuit analog model of CSF system. In this model, current represents flow of CSF fluid, and voltage represents pressure. The capacitance C, represents intracranial compliance, and the resistance R represents resistance of CSF fluid flow into the venous system.

hydrodynamics (125,126). A feature of this model is that the arterial–arteriolar vascular bed consists of vasomotor regulating resistance and compliance elements. Dilatory and constrictive responses are primarily simulated by a corresponding decrease or increase of vascular resistance. This initial model has been modified by synchronously modulating the terminal venous bed resistance to account for cyclic variation of ICP produced by positive pressure ventilation (127). The modified model demonstrates that both the depth of modulation and the cerebrovascular venous terminal bed resistance seem to be progressively reduced with increasing levels of vasodilation induced by increasing the levels of partial pressure of arterial blood carbon dioxide. Numerical modeling of cerebrovascular pressure transmission, the relationship between ABP and CPP, has been used to assess changes in the modes of pressure transmission before and after injury. Specifically, a proposed third-order model of ICP dynamics (128) was used to define the mathematical structure to construct a numerical identification model of cerebrovascular pressure transmission from laboratory pressure recordings (112). Consistent with active vasoconstriction before brain injury, during intact regulation of cerebral blood flow, the highest modal frequency of cerebrovascular pressure transmission decreased with increasing CPP. Conversely, consistent with passive vasodilation with loss of autoregulation induced by fluid percussion injury, the highest modal frequency demonstrated a direct relationship with increasing CPP (112).

SUMMARY

Although the subject of intracranial pressure dates back to over 200 ago, both the invasive monitoring procedure required to measure ICP with its accompanying risk of infection and the nonunique nature of the extravascular pressure measurement have retarded the development of knowledge in this area. As a result, the intent of this article is to provide the reader with a broad perspective of ICP monitoring by providing background material on (1) the physiological and anatomical characteristics of the craniospinal, (2) a history of instrumentation techniques including the most recent advances, (3) a brief review of significant clinical ICP monitoring studies with an emphasis on severe head injury, (4) a review of analysis methods, and (5) models of ICP dynamics related to analysis.

BIBLIOGRAPHY

1. Stern WE. Intracranial fluid dynamics. The relationship of intracranial pressure to the Monro-Kellie doctrine and the reliability of pressure assessment. J Royal College Surgeons Edinburgh 1963;9:18–36.
2. Langfitt TW. Increased intracranial pressure. Clin Neurosurg 1969;6:436–471.
3. Monro A. Observations on the Structure and Function of the Nervous System. Edinburgh: Creech and Johnston; 1783.
4. Kellie G. An account of the appearances observed in the dissection of two of three individuals presumed to have perished in the storm of the third and whose bodies were discovered in the vicinity of Leith on the morning of the 4th, November 1821, with some reflections on the pathology of the brain. Trans Med Chir Soc Edinb 1824;1:84–169.
5. Weed LH, McKibben PS. Pressure changes in cerebrospinal fluid following intravenous injection of solutions of various concentrations. Am J Physiol 1919;48:512–530.
6. Weed LH. Some limitations of the Monro–Kellie hypothesis. Arch Surg 1929;18:1049–1068.
7. Kocher T. Hirnerschutterung, Hirndruck und chirurgische Eingriffe bein Hirnerkrankungen Nothnagel: Spezielle Pathologie und Therapie: 1901.
8. Duret H. Etudes experimentales et cliniques sur les traumatismes cerebrau. Paris: 1978.
9. Cushing H. Some experimental and clinical observations concerning states of increased intracranial tension. Am J Med Sci 1902;124:375–400.
10. Cushing H. The blood pressure reaction of acute cerebral compression, illustrated by cases of intracranial haemorrhage. Am J Med Sci 1903;125:1017–1044.
11. Cushing H. Concerning a definite regulatory mechanism of the vasomotor center which controls blood pressure during cerebral compression. Johns Hopk Hosp Bull 1901;12:290–292.
12. Jennet WB. Experimental brain compression. Arch Neurol 1961;4:599–607.
13. Johnston IH, Rowan JO. Raised intracranial pressure and cerebral blood flow: 3. Venous outflow tract pressures and vascular resistances in experimental intracranial hypertension. J Neurol Neurosurg Psych 1974;37:392–402.
14. Browder J, Meyers R. Observations on behaviour of the systemic blood pressure, pulse and spinal fluid pressure following craniocerebral injury. Am J Surg 1936;31:403–427.
15. Smyth CE, Henderson WR. Observations on the cerebrospinal fluid pressure on simultaneous ventricular and lumbar punctures. J Neurol Psychiat 1938;1:226–237.
16. Evans JP, Espey FF, Kristoff FV, Kimball FD, Ryder HW. Experimental and clinical observations on rising intracranial pressure. Arch Surg 1951;63:107–114.
17. Marmarou A, Shulman K, LaMorgese J. Compartmental analysis of compliance and outflow resistance of the cerebrospinal fluid system. J Neurosurg 1975;43:523–534.
18. Miller JD, Stanek AE, Langfitt TW. Concepts of cerebral perfusion pressure and vascular compression during intracranial hypertension. In: Meyer JS, Schade JP, editors. Progress in Brain Research. vol 35: Cerebral Blood Flow. Amsterdam: Elsevier; 1972. pp 411–432.
19. Miller JD, Garibi J, Pickard JD. Induced changes of cerebrospinal fluid volume: Effects during continuous monitoring of ventricular fluid pressure. Arch Neurol 1973;28: 265–269.
20. Miller JD, Pickard JD. Intracranial volume/pressure studies in patients with head injury. Injury 1974;5:265–268.
21. Miller JD, Leech PJ, Pickard JD. Volume pressure response in various experimental and clinical conditions. In: Lundberg N, Ponten U, Brock M, editors. Berlin: Springer; Intracranial Pressure II; 1975. pp 97–99.
22. Miller JD. Volume and pressure in the craniospinal axis. Clin Neurosurg 1975;22:76–105.
23. Hase U, Reulen HJ, Meinig G, Schrmann K. The influence of the decompressive operation on the intracranial pressure and the pressure volume relation in patients with severe head injuries. Acta Neurochir 1978;45:1–13.
24. Lundberg N. Continuous recording and control of ventricular fluid pressure in neurosurgical practice. Acta Psychiat Neurol Scand 1960;36:149.
25. Baylis WM, Hill L, Gulland GL. On intracranial pressure and the cerebral circulation. J Physiol 1897;18:334–362.
26. Rosner MJ, Becker DP. ICP monitoring: Complications and associated factors. Clin Neurosurg 1996;23:494–519.

27. Maas A, Dearden M, Teasdale GM, Braakman R, Cohodan F, Iannotti F, Karimi A, Lapirre F, Murray G, Ohman J, Persson L, Servadei F, Stocchetti N, Unterburg A. EBIC-Guidelines for management of severe head injury in adults. Acta Neurochir 1997;139:286–294.
28. Naryan RK. et al. Guidelines for the management of severe head injury. J Neurotrauma 1997;13:639–734.
29. Gabe IT. Cardiovascular fluid dynamics. Acta Physiol Scand 1972;19:306–322.
30. Piper IR, Dearden NM, Miller JD. Methodology of spectral analysis of the intracranial pressure waveform in a head injury intensive care unit. In: Hoff JT, Betz AL, editors. Intracranial Pressure VII. Berlin: Springer-Verlag; 1989. pp 668–671.
31. Allan MWB. Measurement of arterial pressure using catheter–transducer systems. Br J Anaesth 1988;60:413–418.
32. Czosnyka M, Czosnyka Z, Pickard J. Laboratory testing of the Spiegelberg brain pressure monitor: A technical report. J Neurol Neurosurg Psych 1997;63:732–735.
33. Ostrup RC, Luerssen TG, Marshall LF. Continuous monitoring of intracranial pressure with a miniturized fibreoptic device. J Neurosurg 1987;67:206–209.
34. Marmarou A, Tsuji O, Dunbar JG. Experimental evaluation of a new solid state ICP monitor. In: Nagai H, Kemiya K, Ishiii S, editors. Intracranial Pressure IX. New York: Springer-Verlag; 1994. pp 15–19.
35. Morgalla MH, Cuno M, Mettenleiter H, Will BE, Krasznai L, Skalej M, Bitzer M, Grote EH. ICP monitoring with a reusable transducer: experimental and clinical evaluation of the Gaeltec ICT/b pressure probe. Acta Neurochir (Wien) 1997;139(6): 569–573.
36. Holzschuh M, Woertgen C, Metz C, Brawanski A. Clinical evaluation of the InnerSpace fibreoptic intracranial pressure monitoring device. Brain 1988;12:191–198.
37. Betsch HM, Aschoff A. Measurement artifacts in Gaeltec intracrnial pressure monitors due to radio waves from personal beeper systems. Anaesthesiol Intensivmed Notfallmed Schmerzther 1992;27(1):51–52.
38. Munch E, Weigel R, Schmiedek P, Schurer L. The Camino intracranial pressure device in clinical practice: Reliability, handling characteristics and complications. Acta Neurochir 1998;140:1113–1119.
39. Chambers IR, Kane PJ, Choksey MS, Mendelow AD. An evaluation of the Camino ventricular bolt system in clinical practice. Neurosurgery 1993;33:866–868.
40. Weinstable C, Richling B, Plainer B, Czech T, Spiss CK. Comparative analysis between epidural (Gaeltec) and subdural (Camino) intracranial pressure probes. J Clin Monit 1992;8:116–120.
41. Piper IR, Miller JD. The evaluation of the waveform analysis capability of a new strain-gauge intracranial pressure microsensor. Neurosurgery 1995;36:1142–1145.
42. Signorini DF, Shad A, Piper IR, Statham PFX. A clinical evaluation of the Codman microsensor for intracranial pressure monitoring. Br J Neurosurgery 1988;12:223–227.
43. Fernades HM, Bingham K, Chambers IR, Mendelow AD. Clinical evaluation of the Codman microsensor intracranial pressure monitoring system. Acta Neurochir Suppl (Wien) 1998;71:44–66.
44. Mendelow AD, Rowan JO, Murray L, Kerr AE. A clinical comparison of subdural screw pressure measurements with ventricular pressure. J Neurosurg 1983;58:45–50.
45. Piper IR, Barnes A, Smith DH, Dunn L. The Camino intracranial pressure sensor: Is it optimal technology? An internal audit with a review of current intracranial pressure monitoring technologies. Neurosurgery 2001;49: 1158–1165.
46. Citerio G, Piper I, Cormio M, Galli D, Cazzaniga S, Enblad P, Nilsson P, Contant C, Chambers I, on behalf of the BrainIT Group, Bench test assessment of the new Raumedic.
47. Piper IR, Miller JD, Dearden NM, Leggate JR, Robertson I. Systems analysis of cerebrovascular pressure transmission: An observational study in head injured patients. J Neurosurg 1993;73:871–880.
48. Becker DP, et al. The outcome from severe head injury with early diagnosis and intensive management. J Neurosurg 1977;47:491–502.
49. Marshall LF, Smith RW, Shapiro HM. The outcome with aggressive treatment in severe head injuries. Part 1: The significance of intracranial pressure monitoring. J Neurosurg 1979;50:20–25.
50. Miller JD, et al. Significance of intracranial hypertension in severe head injury. J Neurosurg 1977;47:503–516.
51. Pitts LH, Kaktis JV, Juster R, Heilbrun D. ICP and outcome in patients with severe head injury. In: Shulman K, Marmarou A, Miller JD, Becker DP, Hochwald GM, Brock M, editors. Intracranial Pressure IV. Berlin: Springer; 1980. pp 5–9.
52. Stuart G, Merry G, Smith JA, Yelland JDM. Severe head injury managed without intracranial pressure monitoring. J Neurosurg 1983;59:601–605.
53. Miller JD, et al. Further experience in the management of severe head injury. J Neurosurg 1981;54:289–299.
54. Alberico AM, Ward JD, Choi SC, Marmarou A, Young H. Outcome after severe head injury: Relationship to mass lesions, diffuse injury and ICP course in pediatric and adult patients. J Neurosurg 1987;67:648–656.
55. Choi SC, Muizelaar JP, Barnes TY, Marmarou A, Brooks DM, Young HF. Prediction tree for severely head injured patients. J Neurosurg 1991;75:251–255.
56. Vollmer DG, Torner JC, Jane J, et al. Age and outcome following traumatic coma: Why do older patients fare worse? J Neurosurg 1991;75:s37–s49.
57. O'Sullivan MG, Statham PF, Jones PA, et al. Role of intracranial pressure monitoring in severely head–injured patients without signs of intracranial hyperension on initial computerised tomography. J Neurosurg 1994;80:46–50.
58. Narayan RK, Greenberg RP, Miller JD, Enas GG, Choi SC, Kishore PRS.
59. Saul TG, Ducker TB. Effect of intracranial pressure monitoring and aggressive treatment on mortality in severe head injury. J Neurosurg 1982;56:498–503.
60. Marmarou A, et al. Impact of ICP instability on outcome in patients with severe head trauma. J Neurosurg 1991;75:s59–s66.
61. Jones PA, Andrews PJD, Midgley S, Anderson SI, Piper IR, Tocher JL, Housley AM, Corrie JA, Slattery J, Dearden NM, Miller JD. Measuring the burden of secondary insults in head–injured patients during intensive care. J Neurosurg Anaesth. 6:4–14.
62. Lewelt W, Jenkins LW, Miller JD. Autoregulation of cerebvral blood flow after experimental fluid percussion injury of the brain. J Neurosurg 1980;53:500–511.
63. Povlishock JT, Kontos HA. Continuing axonal and vascular change following experimental brain trauma. Central Nervous System Trauma 1985;2:285–298.
64. Nordstrom CH, et al. Cerebral blood flow, vasoreactivity and oxygen consumption during barbiturate therapy in severe traumatic brain lesions. J Neurosurg 1988;68:424–431.
65. Miller JD, Adams JH. The pathophysiology of raised intracranial pressure. In: Adams JH, Duchen LW, editors. Greenfield's Neuropathology. 5th ed. London: Arnold; 1992: 69–105.
66. DeSalles AF, Muizelaar JP, Young HF. Hyperglycemia, cerebrospinal fluid lactic acidosis and cerebral blood flow in severly head–injured patients. Neurosurgery 1987;21:45–50.

67. Jaggi LJ, Obrist WD, Gennarelli TA, Langfitt TW. Relationship of early cerebral blood flow and metabolism to outcome in acute head injury. J Neurosurg 1990;72:176–182.
68. Muizelaar JP, Ward JD, Marmarou A, Newton PG, Wachi A. Cerebral blood flow and metabolism in severely head–injured children. Part 2: Autoregulation. J Neurosurg 1989;71:72–76.
69. Muizelaar JP, Marmarou A, DeSalles AAF, Ward JD, Zimmerman RS, Zhongchao L, Choi SC, Young HF. Cerebral blood flow and metabolism in severely head–injured children. Part 1: Relationship with GCS score, outcome, ICP and PVI. J Neurosurg 1989;71:63–71.
70. Uzzell BP, Obrist WD, Dolinskas CA, Langfitt TW. Relationship of acute CBF and ICP findings to neuropsychological outcome in severe head injury. J Neurosurg 1986;65:630–635.
71. Rose J, Valtonen S, Jennett B. Avoidable factors contributing to death after head injury. Brit Med J 1977;2:615–618.
72. Gentleman D, Jennett B. Hazards of interhospital transfer of comatose head injured patients. Lancet 1981;ii:835–855.
73. Graham DI, Hume Adams J, Doyle D. Ischemic brain damage in fatal non–missile head injuries. J Neurol Sci 1978;39:213–234.
74. Graham DI, Ford I, Adams JH, Doyle D, Teasdale GM, Lawrence AE, Mclellan DR. Ischemic brain damage is still common fatal non–missile head injury. 1989.
75. Miller JD, Garibi J. Intracranial volume/pressure relationships during continuous monitoring of ventricular fluid pressure. In: Brock M, Dietz H, editors. Intracranial Pressure. Berlin: Springer; 1972. 270–274.
76. Johnston IH, Rowan JO. Raised intracranial pressure and cerebral blood flow: 3. Venous outflow tract pressures and vascular resistances in experimental intracranial hypertension. J Neurol Neurosurg Psychi 1974;37:392–402.
77. Yada K, Nakagawa Y, Tsuru M. Circulatory disturbance of the venous system during experimental intracranial hypertension. J Neurosurg 1973;39:723–729.
78. Nakagawa Y, Tsura M, Yada K. Site and mechanism for compression of the venous system during experimental intracranial hypertension. J Neurosurg 1974;41:427–434.
79. Harper AM. Autoregulation of cerebral blood flow: influence of the arterial pressure on blood flow through the cerebral cortex. J Neurol Neurosurg Psych 1966;29:398–403.
80. Muizelaar JP, Becker DP. Induced hypertension for the treatment of cerebral ischemia after subarachnoid hemorrhage. Direct effect on cerebral blood flow. Surg Neurol 1986;25:317–325.
81. Obrist WD, et al. Cerebral blood flow and metabolism in comatose patients with acute head injury. Relationship to intracranial hypertension. J Neurosurg 1984;61:241–253.
82. Bouma GJ, Muizelaar JP. Relationship between cardiac output and cerebral blood flow in patients with intact and with impaired autoregulation. J Neurosurg 1990;73:368–374.
83. Chan KH, Miller JD, Piper IR. Cerebral blood flow at constant cerebral perfusion pressure but changing arterial and intracranial pressure: Relationship to autoregulation. J Neurosurg Anaesth 1992;4:188–193.
84. Piper IR, Lawson A, Dearden NM, Miller JD. Computerised data collection: a microcomputer data collection system in head injury intensive care. Br J Intensive Care 1991;1:73–78.
85. Chambers IR, Treadwell L, Mendelow AD. The cause and incidence of secondary insults in severely head–injured adults and children. 2000.
86. Signorini DF, Andrews PJD, Jones PAJ, Wardlaw JM, Miller JD. Prediciting survival using simple clinical variables: A case study in traumatic brain injury. J Neurol Neurosurg Psych 1999;66:20–25.
87. Signorini DF, Andrews PJD, Jones PAJ, Wardlaw JM, Miller JD. Adding insult to injury: The prognostic value of early secondary insults for survival after traumatic brain injury. J Neurol Neurosurg Psychi 1999;66:26–31.
88. Robertson CS, Valadka AB, Hannay HJ, Contant CF, Gopinath SP, Cormio M, Uzura M, Grossman RG. Prevention of secondary ischemic insults after severe head injury. Crit Care Med 1999;27(10):2086–2095.
89. Chambers IR, Jones PA, Lo TY, Forsyth RJ, Fulton B, Andrews PJ, Mendelow AD, Minns RA. critical threshold of intracranial pressure and cerebral perfusion pressure related to age in paediatric head injury. 2005.
90. Piper I, et al. The BrainIT Group: Concept and core dataset definition. Acta Neurochir 2003;145:615–629.
91. Rosner MJ, Rosner SD, Johnson AH. Cerebral perfusion pressure: management protocol and clinical results. J Neurosurg 1995;83:949–962.
92. Bearing EA. Choroid plexus and arterial pulsation of cerebrospinal fluid. Arch Neurol Psych 1955;73:165–172.
93. Hamit HF, Beall AC, DeBakey ME. Hemodynamic influences upon brain and cerebrospinal fluid pulsations and pressures. J Trauma 1965;5:174–184.
94. Dardenne G, Dereymaeker A, Lacheron JM. cerebrospinal fluid pressure and pulsatility. An experimental study of circulatory and respiratory influences in normal and hydrocephalic dogs. 1969.
95. Hamer J, Alberti E, Hoyer S, Wiedemann K. Influence of systemic and cerebral vascular factors on the cerebrospinal fluid pulse waves. 1977.
96. Avezatt CJJ, van Eijndhoven JHM. Clinical observations on the relationship between cerebrospinal fluid pulse pressure and intracranial pressure. In: Cerebrospinal Fluid Pulse Pressure and Cranio–spinal Dynamics: A Theoretical, Clinical and Experimental Study. Thesis. Erasmus Univ., Rotterdam, 1984.
97. Avezatt CJJ, van Eijndhoven JHM. Cerebrospinal fluid pulse pressure and intracranial volume–pressure relationships. J Neurol Neurosurg Psych 1979;42:687–700.
98. Daley ML, Gallo A, Mauch W. Analysis of the intracranial pressure pulsation associated with the cardiac cycle. Innovation Et Technology En Biologie Et Medicine 1986;7:537–544.
99. Czosynka M, Guazzo E, Whitehouse M, Smielewski P, Czosynka Z, Kirkpatric P, Piechnik S, Pickard JD. Significance of intracranial pressure waveform analysis after head–injury. Acta Neurochir (Wien) 1996;138(5):531–541.
100. Balestreri M, Czosynka M, Steiner LA, Schmidt E, Smielewski P, Matta B, Pickard JD. Intracranial hypertension: What additional information can be derived from ICP waveform after head injury? Acta Neurochir 2004;146:131–141.
101. Steiner IA, Balestreri M, Johnston AJ, Coles JP, Smielewski P, Pickard JD, Menon DK, Czosnyka M. Predicting the response of intracranial pressure to moderate hyperventilation Acta Neurochirurgica (online) 2005.
102. Lenfeldt N, Anderson N, Agren–Wilsson A, Bergenheim AT, Koskinen LO, Eklund A, Malm J. Cerebrospinal fluid pulse pressure method: A possible substitute for the examination of B waves. J Neurosurg 2004;101(6):944–950.
103. Daley ML, Pasley R, Connolly M, Angel J, Timmons S, Stidham G, Leffler C. Spectral characteristics of B–waves and other low–frequency activity. Acta Neurochurirgica 2002;81:147–150.
104. Chopp M, Portnoy H. System analysis of intracranial pressure. J Neurosurg 1980;53:516–527.
105. Portnoy HD, Chopp M. Cerebrospinal fluid pulse wave form analysis during hypercapnia and hypoxia. Neurosurgery 1981;9:14–27.
106. Portnoy HD, Chopp M, Branch C, Shannon MD. Cerebrospinal fluid pulse waveform as an indicator of cerebral autoregulation. J Neurosurg 1982;56:666–678.

107. Piper IR, Chan KH, Whittle IR, Miller JD. An experimental study of cerebrovascular resistance, pressure transmission, and craniospinal compliance. Neurosurgery 1993;32:805–816.
108. Nichols JS, Beel JA, Munro LG. Detection of impaired cerebral autoregulation using spectral analysis of intracranial pressure waves. J Neurotrauma 1996;13:439–456.
109. Bruce DA, et al. Diffuse cerebral swelling following head injury in children: The syndrome of “malignant brain edema”. J Neurosurg 1981;54:170–178.
110. Daley ML, Pasupathy H, Griffith M, Robertson JT, Leffler C. Evaluation of autoregulation of cerebral blood flow by correlation of arterial and intracranial pressure signals. IEEE Trans Biomed Eng 1995;42:420–424.
111. Bruce DA, et al. Pathophysiology, treatment and outcome following severe head injury in children. Child’s Brain 1979;5:174–191.
112. Daley ML, Patterson S, Marmarou A, Leffler CW, Stidham G. Pediatric traumatic brain injury: Correlation of intracranial and arterial pressure signals, Proc. 18th Annu. Int. Conf. IEEE Engineering in Medicine and Biology Society, Amsterdam, Netherlands, Nov. 1996.
113. Czosnyka M, Smielewski P, Kirkpatrick P, Laing R, Menon D, Pickard J. Continuous assessment of the cerebral vasomotor reactivity in head injury. Neurosurgery 1997;41:11–19.
114. Czosnyka M, Smielewski P, Kirkpatrick P, Laing R, Menon D, Pickard J. Continuous assessment of the cerebral vasomotor reactivity in head injury. Acta Neurochirugica 1998;71:74–77.
115. Czosnyka M, Smielewski P, Piechnik S, Schmidt E, Al–Rawi PG, Kirkpatrick PJ, Pickard JD. Hemodynamic characterization of intracranial pressure plateau waves in head–injured patients. J Neurosurg 1999;92:11–19.
116. Steiner LA, Czosynka M, Piechnik SK, Smielewski P, Chatfield D, Menon DK, Pickard JD. Continuous monitoring of cerebrovascular pressure reactivity allows determination of optimal cerebral perfusion pressure in patients with traumatic brain injury. Critical Care Med 2002;30: 733–738.
117. Steiner LA, Coles JP, Johnston AJ, Chatfield DA, Smielewske P, Fryer TD, Aigvirhio FI, Clark JC, Pickard JD, Menon DK, Czosynka M. Assessment of cerebrovascular autoregulation in head–injured patients. Stroke 2003;34: 2404–2409.
118. Oertel M, Kelly DF, Lee JH, Glenn TC, Vespa M, Martin NA. Is CPP therapy beneficial for all patients with high ICP. Acta Neurochir 2002;81:67–68.
119. Howells T, Elf K, Jones PA, Ronne–Engstrom E, Piper I, Nilsson P, Andrews P, Enbald P. Pressure reactivity as a guide in the treatment of cerebral perfusion pressure in patients with brain trauma. J Neurosurg 2005;102(2):311–317.
120. Marmarou A. A theoretical and experimental evaluation of the cerebrospinal flid system. Ph.D. Dissertation, Drexel University, 1973.
121. Marmarou A, Shulman K, LaMorgese J. Compartmental analysis of compliance and outflow resistance of the cerebrospinal fluid system. J Neurosurg 1975;43:523–534.
122. Marmarou A, Shulman K, Rosende RM. A nonlinear analysis of the cerebrospinal fluid system and intracranial pressure dynamics. J Neurosurg 1978;48(3):332–344.
123. Ekstedt J. CSF hydrodynamic studies in man. 1. Method of constant pressure CSF infusion. J Neurol Neurosurg Psych 1977;40(2):105–119.
124. Agarwal GC, Berman BM, Stark L. A lumped parameter model of the cerebrospinal fluid system. IEEE Trans Biomed Eng 1969;16:45–53.
125. Ursino M. A mathematical study of human intracranial hydrodynamics. Part 1—the cerebrospinal fluid pulse pressure. Ann Biomed Eng 1988;16:379–401.
126. Ursino M. A mathematical study of human intracranial hydrodynamics. Part 2—Simulation of clinical tests. Ann Biomed Eng 1988;16:403–416.
127. Pasley RL, Leffler CW, Daley ML. Modeling modulation of intracranial pressure by variatrion of cerebral venous resistance induced by ventilation. Ann Biomed Eng 2003;31: 1238–1245.
128. Czosnyka M, Piechnik S, Richards HK, Kirkpatrick P, Smielewski P, Pickard JD. Contribution of mathematical modeling to the interpretation of bedside tests of cerebrovascular autoregulation. J Neurol Neurosurg Psych 1997;63: 721–731.

See also BIOTELEMETRY; HYDROCEPHALUS, TOOLS FOR DIAGNOSIS AND TREATMENT OF; NEONATAL MONITORING; NEUROLOGICAL MONITORS.

MONITORING, NEONATAL.

See NEONATAL MONITORING.

MONITORING, UMBILICAL ARTERY AND VEIN

AHMAD ELSHARYDAH
HAIBO WANG
RANDALL C. CORK
Louisiana State University
Health Center
Department of Anesthesiology
Shreveport, Louisiana

INTRODUCTION

In the neonatal intensive care unit (NICU), monitoring is an integral part of patient care. The primary goal of monitoring is to ensure that early and appropriate intervention can be initiated before to the onset of complications. Monitoring is also a means by which the effect of interventions and therapies may be recorded, evaluated, and controlled. The NICU staff have to deal with a full range of conditions that can arise in the preterm or critically ill neonate, including hypoxemia, hypoglycemia, hypotension, acidosis, and other serious problems. This has led to the evolution and development of several monitors and different sensor-based technologies for use in NICU monitoring including umbilical vessel monitoring. These sensors may provide more accurate and reliable monitoring of neonatal physiological and biochemical changes with a rapid response time. The umbilical vessels may be directly accessed in the first few days of life. An umbilical artery catheter (UAC) may be used for blood pressure monitoring, blood sampling, and fluid or drug infusion. An umbilical vein catheter (UVC) may be used for central venous pressure monitoring, blood sampling, and fluid or drug infusion. Different types of commercially available umbilical catheters are used for these purposes. These catheters differ in their length, size, number of ports, and their material (such as silicone and polyurethane). Blood pressure (BP) monitoring is an important part of neonatal intensive care both for the acutely ill and the convalescing neonate. The most accurate method of measuring BP is by direct intra-arterial recordings, which usually use an umbilical catheter to access the umbilical artery. As blood gas measurement methods and monitors have progressed in adult critical medicine, most of the new techniques and sensors have been used in the NICU by using umbilical artery catheterization. This article addresses the potential benefits of umbilical vessel catheters and associated monitoring devices. It sheds light on the catheters and monitors available on the market and explains the complications and the risks of these catheters. Furthermore, this article looks at the direction of this technology in the future, and it tries to stimulate development of new technology for use in the monitoring of critically ill newborn infants (1–3).

Historical Aspects

In 1946, Louis K. Diamond, a pediatrician from Boston, and F. H. Allen, Jr. developed a technique that allowed blood transfusion to take place through the infant's umbilical cord vein. Regular transfusions were difficult because of the small size of blood vessels in newborns, and there was a further complication due to the use of steel needles and rubber catheters. Diamond used plastic tubing on the umbilical vein, which was larger than average and remained open for several days after birth (4). By the 1960s, electronic monitors came into use and blood gases began to be measured. By the 1970s, the use of umbilical catheters and arterial pressure transducers was routine (5). The first organized NICU opened its doors at Yale-New Haven Hospital in 1960. The first successful use of extracorporeal membrane oxygenation (ECMO) was in 1975. ECMO eventually reduced infant mortality from 80% to 25% for the critically ill infants with acute reversible respiratory and cardiac failure unresponsive to conventional therapy (6).

Anatomical and Physiological Aspects

The umbilical cord is a cordlike structure about 56 cm long, extending from the abdominal wall of the fetus to the placenta. Its chief function is to carry nutrients and oxygen (O_2) from the placenta to the fetus and return waste products and carbon dioxide (CO_2) to the placenta from the fetus. It consists of a continuation of the membrane covering the fetus and encloses a mucoid jelly (Wharton's jelly) with one vein and two arteries (7). Examination of the umbilical cord (after cut) normally reveals two umbilical arteries (UA) and one umbilical vein (UV) (Fig. 1). At skin level, the UV is usually in the 12 o'clock position and has a thinner wall and wider lumen than do the UAs (2). Before birth, blood from the placenta, about 80% saturated with O_2, returns to the fetus by way of the UV. On approaching the liver, most blood flows through the ductus venous directly into the inferior vena cava (IVC), short-circuiting the liver. A smaller amount enters the liver sinusoids and mixes with blood from the portal circulation. After a short course in the IVC, it mixes with deoxygenated blood returning from the lower limbs before it enters the right atrium. The blood leaves the heart to the descending aorta, where it flows toward the placenta by way of the two UAs (7). The O_2 saturation in the umbilical arteries is approximately 58%. Changes in the vascular system at birth are caused by cessation of placental blood flow and the beginning of respiration. These changes are summarized in closure of the umbilical arteries, closure of the umbilical vein and ductus venous, significant reduction in the pulmonary vascular resistance and right ventricle and right atrium pressures, and significant increase in the systemic vascular resistance, left ventricle, and left atrium pressures (8). After birth, the blood volume of the neonate is about 300 mL, the cardiac output averages 500 mL/min, and the arterial blood pressure is about 70/50 during the first day, which increases slowly over the next several months. Arterial blood pressure of the neonate has been best correlated with birth weight. Moreover, systolic and diastolic pressures in the neonate are significantly correlated to blood pressure levels in the mother (9).

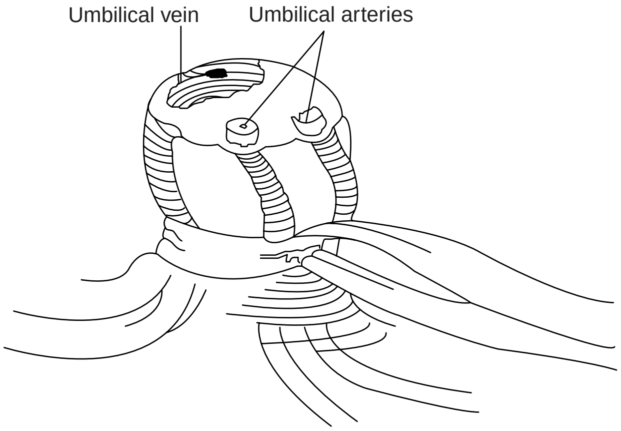


Figure 1. Umbilical cord after it was cut.

Indications and Contra-Indications for Umbilical Artery Catheterization

The primary indications for umbilical artery catheterization include frequent or continuous blood gas measurements, continuous monitoring of arterial blood pressure, and resuscitation (umbilical venous line may be the first choice) (10). Secondary indications include infusion of maintenance glucose-electrolyte solutions or medications, exchange transfusions, angiography, and a port for frequent blood sampling, especially in a very low-birth-weight neonate. These catheters should stay in place only as long as a primary indication exists, with the exception of the very low-birth-weight neonate who may need it for vital infusions and frequent blood sampling. Contraindications include evidence of local vascular compromise in lower limbs or buttock areas, peritonitis, necrotizing enterocolitis, omphalitis, gastroischisis, and omphalocele.

Indications and Contra-Indications for Umbilical Vein Catheterization

The most frequent indications for umbilical venous catheterization include emergency medication administration, exchange transfusion, and partial exchange transfusion (10). This catheter is also used for frequent blood sampling and central venous pressure monitoring. Contraindications for the this catheter include routine fluid infusion

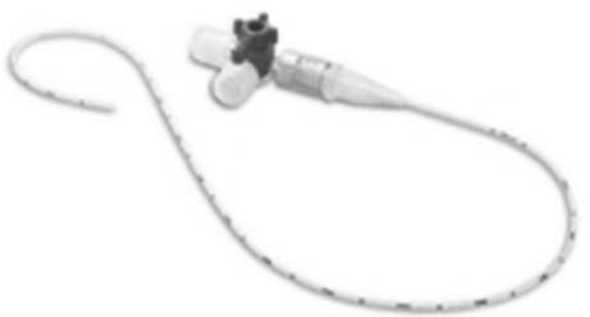


Figure 2. Commercially available umbilical catheters: (a) single-lumen catheter and (b) dual-lumen catheter.

(relative contraindication), omphalitis, omphalocele, gastroischisis, necrotizing enterocolitis, peritonitis, and extrophy of the bladder.

Description of Available Catheters (Design and Material)

Catheters must be made of nontoxic materials that are the least injurious to the vascular intima and least likely to cause thromboses and early atherosclerotic lesions (11). Commercially available umbilical catheters are usually latex-free, made of silicon or high-quality aliphatic polyurethane elastomer (Tecoflex). Silicone catheters are soft, not-irritating, usually not reactive to body tissues and body fluids, not supportive of bacterial growth, and less problematic with blood clotting. Tecoflex is an advanced medical formulation of polyurethane. Its physical characteristics are very close to silicone; however, it is slightly stiffer during insertion, which makes it easier to insert and better to conduct arterial pressure. Tecoflex is thermosensitive, softens at body temperature (12), and significantly reduces the trauma to the vascular intima. These catheters have a rounded tip, which makes them less likely to perforate through the vessel during insertion. They have depth markings at every centimeter for more accurate placement and encased radiopaque stripes to confirm placement by X ray after placement. Umbilical catheters are available with single-lumen, dual-lumen, and triple-lumen catheters (13) (Fig. 2,a,b) in different sizes ranging from 2.6 to 8.0 Fr. (Table 1) Umbilical artery catheter tips are designed in two different ways: end-hole and side-hole tips. The side-hole catheters use a special electrode to measure blood gases and biochemicals. A Cochrane review by Barrington (14) showed that end-hole catheters are associated with a much decreased risk of aortic thrombosis compared with side-hole catheters. Therefore, umbilical artery catheters designed with a side-hole should not be used routinely for umbilical artery catheterization in the newborn. Furthermore, manufactures have made sterile, ready-to-go umbilical catheterization trays. These trays contain everything needed for umbilical catheterization, including drapes, towels, suture, umbilical tape, skin preparation materials, needles, forceps, syringes, and other instruments.

Table 1. Neonate Weight and Umbilical Catheter Size

Neonate Weight (g)	UAC Size (Fr)	UVC Size (Fr)
<1500 g	3.5 Fr	3.5 Fr
>1500 g	5.0 Fr	5.0 Fr

Umbilical Vessel Catheterization Procedure

The infant must be supine and restrained (2). A field around the umbilicus is sterilized and draped, and a silk suture is looped around the base of the umbilical stump. The distal end of the stump is cut off, leaving 2 cm of stump, and the vessels are occluded to prevent blood loss. For the umbilical artery catheterization (15), the stump is firmly grasped with the gloved fingers of one hand, and one of the two thick-walled umbilical arteries is dilated with a curved iris forceps; then the umbilical artery catheter is inserted into the artery. Some resistance may be encountered when the catheter has been advanced 3 to 5 cm into the vessel, but this resistance can usually be overcome by applying steady downward pressure on the catheter. If the catheter cannot advance, a second catheter can be inserted into the other artery while leaving the first catheter in place. This maneuver often causes one or the other vessel to relax and permits one catheter to be advanced into the aorta. Advancement of an UAC should place the tip above the celiac axis but below the ductus arteriosis. All air should be removed from the system. The accidental injection of small amounts of air (<0.1 mL) may obstruct blood flow to the legs for several hours. The catheter should be attached to a pressure transducer and the arterial pressure measured. For the umbilical vein catheterization, the single, large, thin-walled umbilical vein is grasped with an iris forceps, and the air-free catheter, which is connected to a closed stopcock, is inserted 3 to 5 cm into the vessel with a twisting motion. The UVC tip should lie a few centimeters into the umbilical vein or inferior vena cava. The stopcock must be closed to prevent aspiration of air through the catheter should the patient take a deep breath. It is imperative that no air be injected through venous catheter, because the air may enter the systemic circulation through the foramen ovale and occlude a coronary or cerebral artery. If it does, the neonate may die or suffer central nervous system damage. If the catheter "tickles" the atrial septum, the neonate may suffer arrhythmias. Withdrawal of the catheter a short distance can solve the problem. Plain radiographs should be taken to confirm placement (Fig. 3). "High" placement of the UAC is defined in one major review of the literature as one with "the tip in the descending aorta above the level of the diaphragm and below the left subclavian artery" and "low" placement of the UAC as one with "the tip above the aortic bifurcation and below the renal arteries" (16).

Neonatal Blood Gas and Biochemical Measurement

The blood gas measurement is the most widely used clinical method for assessing pulmonary function in the neonate. It forms the basis for diagnosis and management of neonates with cardiorespiratory disease (17). The physiology of blood gases is discussed in other parts of this Encyclopedia. We will discuss in this section some issues related to the neonatal blood gas measurement. The dissociation curve of fetal hemoglobin (as compared with adult) is shifted to

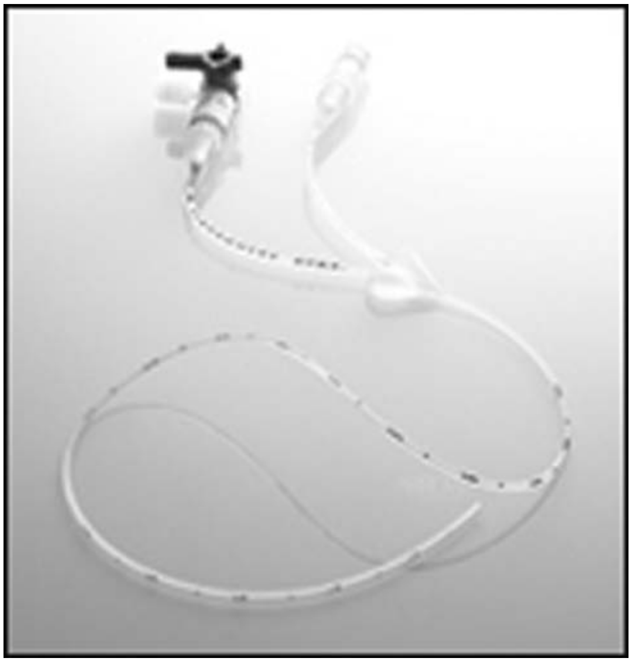


Figure 3. Anteroposterior roentgenogram shows the position of umbilical artery and vein catheters. Lateral roentgenogram is needed to distinguish the umbilical artery from the umbilical vein catheter and to determine the appropriate level of insertion. A = endotracheal tube; B = umbilical venous catheter. C = umbilical artery catheter passed up the aorta to T12.

the left, and at any arterial O_2 partial pressure (PaO_2) below 100 mm Hg (13332.2 Pa) fetal blood binds more O_2. This shift seems to be the result of the lower affinity of fetal hemoglobin for 2,3-diphosphoglycerate (DPG). Shunting is a common occurrence in the neonate, such as in congenital cyanotic heart disease, persistent fetal circulation, or atelectasis. O_2 supplementation does not prevent the hypoxia produced by such a shunt. Arterial carbon dioxide partial pressure ($PaCO_2$) is an important measure of pulmonary function in neonatal respiratory disease. The initiation of ventilation with the first breath after normal delivery results in a rapid fall in $PaCO_2$ within minutes of birth. PO_2 rises rapidly to levels of 60 to 90 mm Hg (17). Immaturity of the kidney in the newborn affects the basal acid–base status and the response to additional acid and alkali loads (18). The blood bicarbonate (HCO_{3-}) concentration is typically lower than in the adult (18–21 mEq/L). However, the blood pH (7.35 –7.43) is only marginally decreased because of the compensatory increase in the neonatal respiratory rate. Table 2 lists the normal values of pH, $PaCO_2$, and total CO_2 in the adult and in preterm and term neonates (19). The transition from fetal to neonatal life, which is associated with rapid changes in fluid and electrolyte balance, and the neonate's small size make the electrolytes and glucose assessment difficult and complicated (20), especially in the first week of life. A physiologic decrease in extracellular water volume, as well as a transient increase in serum potassium and transient decreases in plasma glucose and total plasma ionized calcium concentrations, must be taken into account when monitoring neonatal electrolytes and glucose. Frequent and even continuous monitoring of physiological parameters is indicated in some cases, including glucose and calcium monitoring in the premature newborn (limited hepatic glycogen storage) and in newborns of diabetic mothers (21). Before birth, fetal glucose is slightly higher than maternal glucose. With cord clamping, neonatal plasma level plummets over the first 60–90 min of life (23), (23). Neonatal hormonal changes later leads to an increase in endogenous glucose production and stabilization of its level.

Table 2. Acid–base Parameters in Neonates and Adults (mean ± SD)

	Preterm	Term	Adult
pH	7.40 ± 0.08	7.40 ± 0.06	7.40 ± 0.03
*P*CO2	34.0 ± 9.0	33.5 ± 3.6	39.0 ± 2.6
Total CO2	21.0 ± 2.0	21.0 ± 1.8	25.2 ± 2.8

Technical Aspects of Neonatal Blood Gas and Biochemical Measurement

Technologic innovations in the development of biosensors and microprocessors have led to development of bedside small and accurate point-of-care (POC) devices (24). POC devices have been used widely in critical care units, including the NICU. These devices are divided into two groups: "Analyzers," which are not attached to the patient blood source and require blood sampling, and "monitors," which are continuous or near-continuous patient-attached POC monitors (25). Neonatal biochemical measurement may include several important blood parameters, such as sodium, potassium, calcium, glucose, and even lactate Table 3.

Intermittent Blood Gas and Biochemical Measurement (Sampling). This is the most common technique used in the NICU for invasive neonatal blood gas and biochemical measurement. Usually a small blood sample is withdrawn from a blood vessel, such as an umbilical vessel through an umbilical catheter. This sample is analyzed by using a bedside point-of-care analyzer or sent to a small satellite laboratory unit in the NICU or to the central laboratory. Blood gas analyzers are discussed in other articles in this Encyclopedia. These analyzers use the principles of Clark's electrode for PO_2 measurement and Severinghaus's electrode for PCO_2 measurement. Some blood-sample collecting systems are commercially available, such as the Edward VAMP Jr. system manufactured by Edward (Fig. 4), which are manufactured specifically for neonate and small children use. This system is latex-free, disposable, closed, with small volume systems. Such systems are designed to the decrease the risks of blood loss, infection, and air bubbles (26).

Continuous Intravascular Neonatal Blood Gas and Biochemical Sensors. There are several drawbacks for using frequent arterial blood gas (ABG) sampling in the neonate (27). This method may result in blood loss that can necessitate blood transfusion. Moreover, in this method, rapid changes in blood gas values may be missed, especially

Table 3. The Main Blood-Chemistry Parameters Monitored in the Neonatal Care Unit, With Typical Sensing Principles and Transducers

Parameter	Sensing Principle(s)	Transducer(s)
Invasive Blood Pressure	Electrical, impedance	Strain gauge, piezoresistor
	Optical, reflection	Photodetector and emitter
PO_2	Optical, fluorescent	Photomultipler tube
	Electrochemical, amperometric	Clark oxygen electrode
PCO_2	Optical, fluorescent	Photomultipler tube
	Electrochemical, potentiometric	Ion-sensitive electrode
Glucose	Optical, colorimetric	Photodetector
	Electrochemical, amperometric	Enzyme modified biosensor
Lactate	Optical, colorimetric	Photodetector
	Electrochemical, amperometric	Enzyme modified biosensor
Electrolytes (K, Na, Ca, Cl)	Optical, colorimetric	Photodetector
	Electrochemical, potentiometric	Ion-selective electrode (ISE)
pH	Optical, colorimetric	Photodetector
	Electrochemical, potentiometric	Ion-sensitive electrode
Hemoglobin	Optical, absorption	Photodetector and emitters

in conditions needing quick and close ABG monitoring, such as after surfactant administration (28) and during high-frequency ventilation (29). These drawbacks dictate the need for a more efficient real-time way to monitor ABGs (30). For the last two decades, intra-arterial PaO_2 monitoring has been available with the use of a Clark electrode (31) or a multiparameter sensor with an umbilical artery catheter (32). New fiber-optic continuous blood gas monitoring sensors have been validated and used in the neonate with an UAC, such as Neotrend. These devices promise to be safe, easy to use, and accurate in newborns. However, the cost-effectiveness of these devices is still not well established (33) (refer to the blood gas measurement article in this Encyclopedia for details about Neotrend). The *ex vivo* in-line VIA Low Volume Mode blood gas and chemistry monitoring system (VIA LVM Monitor; Metracor Technologies, Inc., San Diego, CA) is an in-line, low-volume POC monitor for neonates and children. Studies have shown promising results in using this monitor in the neonate. However, its cost-effectiveness has not been established yet (25,34). This device measures pH, $PaCO_2$, PO_2, Na^+, K^+, and hematocrit (Hct) by automatically drawing blood (almost 1.5 mL) from a patient's arterial catheter, analyzing it, and reinfusing the blood sample back into the patient. Results are usually displayed in 1–2 min. The operator performs an initial calibration, and then the device performs self-calibration after each sample and at least every 30 min. This machine is compatible with all sizes of UACs and peripheral arterial catheters. Figure 5 shows a diagram of the VIA LVM monitor and its components at the neonate bedside.

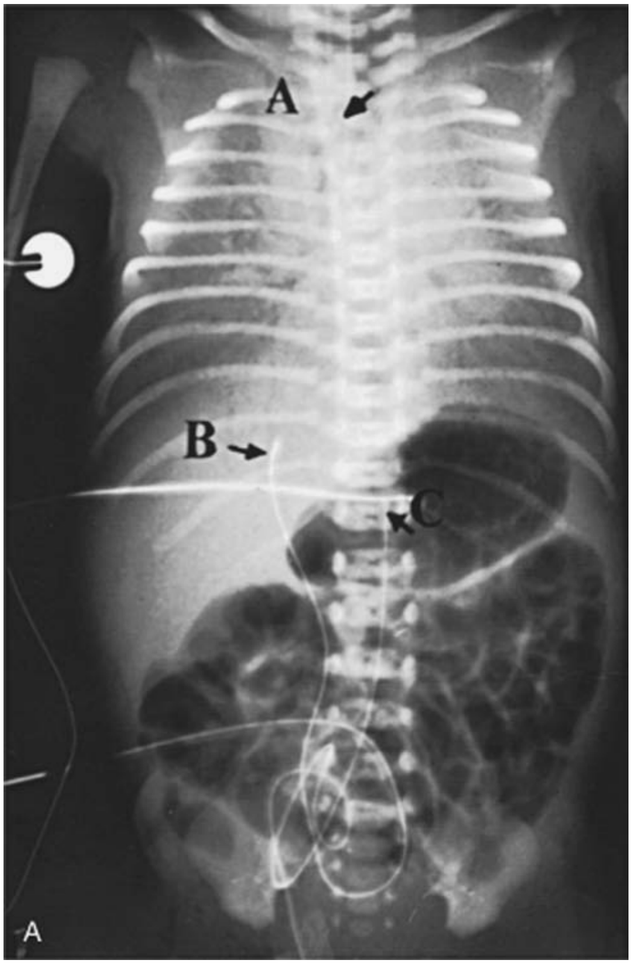


Figure 4. Edward VAMP Jr. blood-sample collecting system.

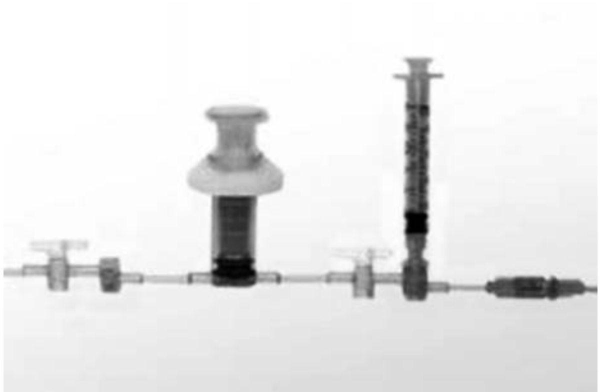


Figure 5. Diagram of the VIA LVM in-line *ex vivo* monitor and its components at the neonate bedside.

Neonatal Hemodynamic Monitoring

Direct arterial blood pressure monitoring is the most accurate technique for determining arterial pressure in the neonate (9). This method is best done by using umbilical artery catheterization. It is an easy and quick procedure in comparison with other neonatal artery catheterizations, such as radial and femoral artery catheterization. After the umbilical artery catheterization is done as described above, it is connected to a pressure transducer and a continuous flow device, as well as stopcock and manometer tubing.

Basic Concepts. Hemodynamic pressure monitoring requires several basic components to accurately measure the physiologic pressures. These components are: (1) an intravascular catheter, (2) connecting tubing and stopcocks to connect that catheter and the patient's blood vessels to the monitoring system, (3) a pressure transducer to convert the mechanical impulse of a pressure wave into an electrical signal through movement of a displaceable sensing diaphragm, (4) a continuous flush device that fills the pressure tubing with fluid and helps prevent blood from clotting in the catheter, (5) an amplifier that increases the low-voltage signal from the pressure transducer to a signal that can be displayed on a display device, (6) an oscilloscope to display waveforms and a digital readout to display numerical data, and (7) a processor or microcomputer that is used to calculate various hemodynamic parameters based on the measured variables.

Pressure Transducers. Pressure transducers are divided in two groups: (1) External transducers located away from the intravascular catheter and connected to that catheter via fluid-filled pressure tubing, and (2) catheter-tip transducers. The external transducers use three types of sensing elements: (1) strain gauges. These consist of an electrically conductive elastic material that responds reversibly to deformation by a change in electrical resistance. The resistance is converted into a voltage signal by connecting the elements to form a Wheatstone bridge circuit. The output voltage is proportional to the applied pressure and the excitation voltage. Strain gauges are the most common method of pressure transduction. (2) Silicon strain gauges. These are thin slices of silicon crystal bonded onto the back of a diaphragm. The movement of the diaphragm causes a change in the resistance of the crystal, which can be converted into an output signal. Silicon strain gauges are more sensitive than standard strain gauges, but they are affected by temperature and are non-linear. (3) Optical sensors: These are also diaphragms, but in this case, the movement of the diaphragm is sensed by reflecting a beam of light off the silver back of the diaphragm onto a photoelectric cell. The intensity of light sensed by the photoelectric cell changes with the diaphragm position, causing a decrease in its electrical output.

The Pressure Measurement System. The arterial waveform can be characterized as a complex sine wave, which is the summation of a series of simple sine waves of different amplitude and frequencies. The fundamental frequency (or first harmonic) is equal to the heart rate. The first 10 harmonics of the fundamental frequency contribute to the waveform. Any measurement system responds to a restricted range of frequencies only. Within this range, the system may respond more sensitively to some frequencies than to others. The response of the system plotted against the signal frequency is the *frequency response* of the system. However, the measurement system may possess *natural frequencies* or resonances determined by the inertial and compliant elements in a mechanical system. These resonances can distort the output signals. Therefore, it is essential that the natural frequencies do not lie in the operating frequency range of the instrument. Moreover, the output signal of a measurement may differ from the input signal created by the bloodstream because of the inertial components, frictional effects of movement, viscous forces of fluids, and electrical resistance. The property that determines these effects is called the *damping* of the system (35).

Technical Management of Pressure Monitoring System. These are some significant practical points in operating this system:

1. *Removal of all air bubbles from system*: Air is more compressible than fluid, and it tends to act as a "shock absorber" within the pressure monitoring system, leading to an overdamped waveform, which may lead to false readings of the blood pressure. Moreover, air bubbles may cause serious air embolism, especially in neonates and small children.
2. *Zeroing the transducer*: The accuracy of invasive pressure measurements is dependent on the establishment of an accurate reference point (Zeroing). This is done by opening the stopcock to atmospheric pressure and zeroing the measurement system to eliminate the effect of the atmospheric pressure, and by leveling the transducer to the level of the upper portion of the right atrium (the patient's "midaxillary line" or "phlebostatic axis") to eliminate the effect of the blood hydrostatic pressure.
3. *Fast-flush technique*: A "fast-flush" or "square wave test" is performed by opening the valve of the continuous flush device, which leads to an acute increase in the fluid flow rate through the catheter-tubing system from the usual 1–3 mL/h to 30 mL/h. This generates an acute rise in pressure within the system such that a square wave is generated on the monitor. With closure of the valve, a sinusoidal pressure wave of a given frequency and progressively decreasing amplitude is generated. A system with appropriate dynamic response characteristics will return to the baseline pressure waveform within one to two oscillations. If the fast-flush technique produces dynamic response characteristics that are inadequate, the clinician should troubleshoot the system (i.e., remove all air bubbles, minimize tubing length and stopcocks, etc.) until an acceptable dynamic response is achieved. The above-explained basic concepts in hemodynamic monitoring are used in pressure monitoring in neonates as well as in adults. Because of

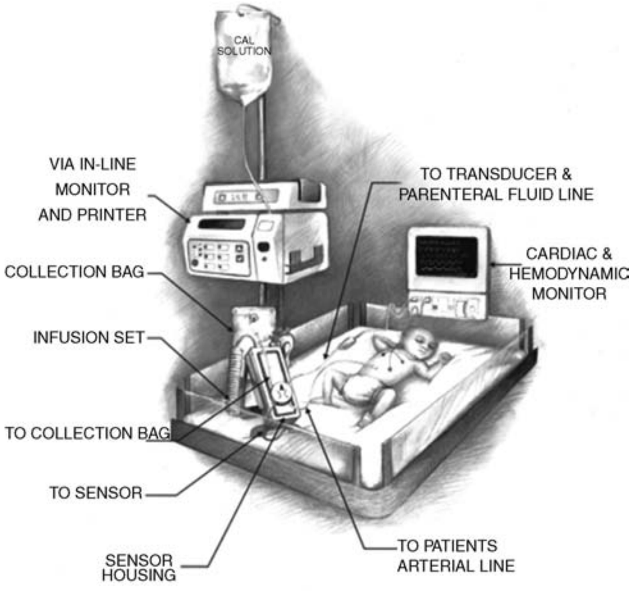

CAL SOLUTION
VIA IN-LINE MONITOR AND PRINTER
TO TRANSDUCER & PARENTERAL FLUID LINE
CARDIAC & HEMODYNAMIC MONITOR
COLLECTION BAG
INFUSION SET
TO COLLECTION BAG
TO SENSOR
SENSOR HOUSING
TO PATIENTS ARTERIAL LINE

Figure 6. Neonatal/Pediatric Deltran pressure tubing with needleless blood collection system.

the small size of a neonate (or premature newborn) and because of the different indications and expected complications from adults and older children, these monitors have been modified to fit these requirements. These modifications and changes include (1) using a more simple tubing system with fewer stopcocks to reduce the risk of air embolism and infection, (2) using a smaller volume tubing system with small syringes to decrease blood loss and the need for blood transfusion, and (3) using a special constant-low-rate flush system to decrease the risk of fluid overloading. After correct placement of the umbilical artery catheter, a stopcock (free of air bubbles) is connected to its distal end. Then a fluid-filled, well-flushed pediatric/neonatal blood pressure tubing is connected to that stopcock (or directly to the catheter). The transducer is zeroed and leveled to the midaxillary level. There are several commercially available neonatal/pediatric pressure transducers and tubing systems. Figure 6 shows a disposable, latex-free neonatal/pediatric Deltran blood pressure monitoring and needleless blood collection system (36). For central venous pressure (CVP) monitoring, a dual-lumen or triple-lumen umbilical vein catheter may be used. Figs. 2 and 7 show some of the currently available catheters. Neonatal/pediatric pulmonary artery catheters (PACs) have been used through the umbilical vein. However, they are difficult to place and have numerous complications and risks associated with their placement. Other noninvasive cardiac monitors, such as echocardiography, can assess cardiac output and other physiological cardiac parameters. Thus, use of the neonatal PAC has decreased significantly. The most common indications for pediatric PACs are for cardiogenic shock, for severe distributive shock, for the use of very high ventilator pressures to achieve adequate oxygenation, and for the perioperative management of patients who have undergone complex cardiac or other major surgeries (37). The smallest thermodilution catheter available is 5 Fr in size, although a single-lumen 4-Fr catheter exists and is useful for measuring pulmonary artery pressure (PAP) or obtaining mixed venous saturation (MvO_2). Potential complications of pulmonary artery catheterization include pulmonary artery erosion or infarction, dysrhythmia, damage to the pulmonic valve, coiling in the right ventricle, and cardiac perforation (38).

Figure 7. Catheter.

Complications and Risks

Umbilical Artery Catheter. Many complications and risks are associated with UAC placement (39). Therefore, these catheters should not be used solely for fluid and medication administration. If an infant does not require frequent arterial blood sampling or continuous blood pressure monitoring, there is almost no justification for leaving a UAC inserted. The advantages and disadvantages of "high" versus "low" UAC are still debated. Recent studies (40–43) have found that "high" catheters are associated with a decreased incidence of complications without a statistically significant increase in any adverse sequelae. Therefore, one major review of the literature has concluded that "there appears to be no evidence to support the use of low placed umbilical artery catheters. High catheters should be used exclusively" (40). It is clear that UACs that are located between the "high" and "low" positions are never appropriate. Catheters in these positions have been associated with refractory hypoglycemia (infusion into the celiac axis), paraplegia (infusion into the artery of Adamkievicz)

Table 4. Care of Indwelling Umbilical Catheter

- Change all tubings and connections daily.
- Secure and label all tubing, and connections should be secured and labeled appropriately.
- Use only appropriate filters.
- Maintain catheter, connections, and tubing free of blood to prevent clot formation, or inadvertent flushing of preexisting clots into the neonate.
- Flush catheter with 0.5 mL of flush solution each time blood sample is drawn.
- Chart fluids infused in intake/output record.
- Infuse heparinized parenteral solution continuously through catheter, interrupting only to obtain blood samples.

(44), and thromboses that affect the kidneys (infusion into the renal arteries) or the gut (infusion into the mesenteric arteries). A catheter that is found in this intermediate position should be pulled to a "low" position or removed. Similarly, catheters should not be placed below the level of L5 because of the risk of gluteal skin necrosis (45) and sciatic nerve damage (45). Catheters that are placed below the level of L5 should be removed promptly. Other complications may include catheter occlusion, infection, air embolism, breaks or transection of the catheter, electrical hazards, intravascular knots in the catheters, bladder injury, peritoneal perforation, Wharton's jelly embolus, and others.

Umbilical Venous Catheter. Thromboembolic events and infections are common complications of UVC use. These complications are similar to those of other central catheters, although UVCs are associated with an increased risk of localized infections of the liver and heart. Complications that are specific to UVC commonly are the result of malposition of the catheter. Nearly all experts recommend placement of the catheter outside the heart in the IVC (46). Complications occur as a result of placement of the catheter in the right side of the heart or the left side (via the foramen ovale). Cardiac arrhythmias are common complications, but these arrhythmias usually resolve after catheter withdrawal from the heart. Cardiac perforation with subsequent pericardial effusion and cardiac tamponade has been reported (47) Quick diagnosis (high index of suspicion, chest radiograph, and ultrasonography) and prompt treatment with pericardiocentesis decrease mortality significantly in these neonates. Placement of the catheter in the portal system can result in serious hepatic injury. Hepatic necrosis can occur from thrombosis of the hepatic veins or infusion of hypertonic or vasospastic solutions into the liver. Necrotizing enterocolitis and perforation of the colon also have been reported after positioning of the catheter in the portal system. Other complications of umbilical venous catheters have been reported and include perforation of the peritoneum, electrical hazards, and digital ischemia.

Umbilical Catheter Removal and Maintenance

According to the American Academy of Pediatrics (AAP) guidelines (48), umbilical artery or venous catheters should be removed and not replaced if there is any sign of catheter-related bloodstream infections, vascular insufficiency, or thrombosis. Moreover, umbilical catheters must be removed as soon as possible when no longer needed or when any sign of vascular insufficiency to the lower extremities is observed. An umbilical artery catheter should not be left in place more than 5 days. However, an umbilical venous catheter can be used up to 14 days if managed aseptically. Umbilical venous catheters may replaced only if the catheter malfunctions. Table 4 lists the important tasks for daily care of umbilical indwelling catheters.

FUTURE TRENDS IN INVASIVE NEONATAL MONITORING

Providing real-time, accurate, reliable, compact, at the bedside, and safe neonatal monitoring devices is the future goal of researchers and manufactures involved with neonatal critical care (3). However, the cost issue is still outstanding and is going to be a major factor in directing this industry. The market demand for integrated critical care units and the clinical demand for continuous and rapid biochemical monitoring at the bedside, especially in critically ill patients, will lead to a closer integration between the vital signs monitors and POC analyzers. The monitor and POC analyzer manufacturers have been separate companies. This will slowly change through the formation of strategic alliances and mergers. Therefore, new devices will combine continuous vital signs, blood gases, and blood chemistry in the critical care units, including the NICU. The demand for continuous monitoring of biochemical parameters is at the same time bounded by the requirement for low-blood-loss systems, especially in the NICU. This is where the in-line and indwelling analyzers can offer a major advantage over the POC analyzers. Fortunately, technology is also moving in the right direction to minimize iatrogenic blood loss and decrease the risk of infection through the advances made in sampling, preparation, and handling of liquids using microfluidic techniques and closed systems (49). It is likely that in-line monitors that interface to umbilical artery catheters will become more widespread and will require decreasing sample-volumes. The continuing application and refinement of established optical assay methods, such as absorption and fluorescence spectroscopy, onto fibre-optic cables will enable the detection of increasing numbers of analytes by indwelling probes encased within the umbilical catheters (50,51).

SUMMARY

The use of umbilical vessel catheterization and associated monitoring techniques and devices has been advanced dramatically since Dr. Louis K. Diamond of Boston used plastic tubing on the umbilical vein for blood transfusion in 1946. Advances in invasive neonatal monitoring and neonatal intravascular access have led to a significant reduction in the incidence of complications and have increased the number of indications for umbilical catheters. Umbilical vessel catheterization is now a routine and safe procedure in the NICU. Moreover, with the increase in the survival rate of low- and very low-birth-weight neonates, the need for these catheters has increased. Sometimes these catheters are the only intravascular access that can be established in this new group of patients. The innovation and development of medical applications of silicone and polyurethane enable the manufacturers to make soft, small catheters with adequate lumen size. These catheters cause minimal adverse reaction to the newborn body. To decrease the risk of blood loss and infection, new blood sampling devices have been developed. These systems are closed and have small volume tubing. New technologies for continuous monitoring of blood gases and neonatal chemistries in-line have been integrated with the umbilical catheters by using special sensors encased in the tip of the catheters. The future trend is to use smaller, compact, real-time, bedside monitors combined with pulse oximetry and other vital signs monitors. However, the issue of the cost-effectiveness of these machines has yet to be determined.

BIBLIOGRAPHY

1. Walsh-Sukys MC, Fanaroff AA. Perinatal services and resources. In: Fanaroff AA, Martin RJ, editors. Neonatal-Perinatal Medicine, Diseases of the Fetus and Infant. 6th ed. St. Louis: Mosby; 1999.
2. Stovroff M, Teague WG. Pediatric surgery for the primary care pediatrician, Part II: Intravenous access in infants and children. Pediatric Clin North Am 1998;45(6).
3. Murkovíc I, Steinberg MD, Murkovíc B. Sensors in neonatal monitoring: Current practice and future trends. Technol Health Care 2003;11:399–412.
4. Diamond LK, Allen Jr FH, Thomas Jr WO, Erythroblastosis fetalis. VII. Treatment with exchange transfusion. N Engl J Med 1951;244:39–49.
5. Klaus MH, Fanaroff AA. Care of the High-Risk Neonate, 5th ed. Philadelphia: WB Saunders; 2001
6. Bartlett RH, Roloff DW, Cornell RG, Andrews AF, Dillon PW, Zwischenberger JB. Extracorporeal circulation in neonatal respiratory failure: A prospective randomized study. Pediatrics 1985;76:479–497.
7. Sadler TW. Langman's Medical Embryology: Cardiovascular System, 8th ed. Philadelphia: Lippincott Williams & Wilkins; 2000.
8. Guyton AC, Hall JE. Textbook of Medical Physiology: Fetal and Neonatal Physiology, 10th ed. Philadelphia: W.B. Saunders; 2000.
9. Zahka KG. Principles of neonatal cardiovascular hemodynamics. In: Fanaroff AA, Martin RJ. editors, Neonatal-Perinatal Medicine, Diseases of the Fetus and Infant. 6th ed. St. Louis: Mosby; 1999.
10. Grady M, Procedures. In: Gunn VL. Nechyba C, editors, Harriet Lane Handbook: A Manual for Pediatric House Officers, 16th ed. St. Louis: Mosby; 2002.
11. Brown EG, Krouskop RW. Monitoring, umbilical artery and vein. In: Webster JG, editor. Encyclopedia of Medical Devices and Instrumentation, New York: Wiley; 1988
12. Chidi CC, King DR, Boles Jr ET, An ultrastructural study of the intimal injury induced by an indwelling umbilical artery catheter. J Pediatr Surg 1983;18:109–115.
13. http://www.utahmed.com/umbili-c.htm.
14. Barrington KJ. Umbilical artery catheters: Catheter design (Cochrane Review). In: The Cochrane Library Issue 4. Oxford; 1997.
15. Gregory GA. Resuscitation of the newborn. In: Miller RD, editor. Miller's Anesthesia. 6th ed. New York: Elsevier; 2005.
16. Barrington KJ. Umbilical artery catheters in the newborn: effects of position of the catheter tip. Cochrane Database Syst Rev CD000505, 2000.
17. Carlo WA. Assessment of pulmonary function. In: Fanaroff AA, Martin RJ, editors. Neonatal-Perinatal Medicine, Diseases of the Fetus and Infant. 6th ed. St Louis: Mosby; 1999.
18. Stork JE, Stork EK. Acid-base physiology and disorders in the neonate. In: Fanaroff AA, Martin RJ. editors, Neonatal-Perinatal Medicine, Diseases of the Fetus and Infant. 6th ed. St. Louis: Mosby; 1999.
19. Lorenz JM, Kleiman LI, Kotagal UR, Reller MD. Water balance in very low birth infants: Relationship to water and sodium intake and effect on outcome J Pediatrics 1982;101: 423–432.
20. Lorenz JM. Assessing fluid and electrolyte status in the newborn, Nat Acad of Clin Biochem. Clin Chem 1997;43(1):205–210.
21. Heck LI, Erenberg A. Serum glucose values during the first 48 h of life. J Pediatr 1978;110:119–122.
22. Srinivasan G, Pildes RS, Cattamanchi G, Voora S, Lillian LD. Plasma glucose values in normal neonates: A new look. J Pediatr 1984;105:114–119.
23. Tsang RC, Chen IW, Freidman MA, Chen I. Neonatal parathyroid function: role of gestational and postnatal age. J Pediatr 1973;83:728–730.
24. Ehrmeyer SS, Laessig RH, Leinweber JE, Oryall JJ. Medicare/CLIA final rules for proficiency testing: Minimum intralaboratory performance characteristics (CV and bias) needed to pass. Clin Chem 1990;36:1736–1740.
25. Billman GF, et al. Clinical performance of an in-line, ex vivo point-of-care monitor: A multicenter study. Clin Chem 2002;48(11): 2030–2043.
26. http://www.edwards.com/Products/PressureMonitoring/Vamp Jr.htm
27. Meyers PA, Worwa C, Trusty R, Mammel MC. Clinical validation of a continuous intravascular neonatal blood gas sensor introduced through an umbilical artery catheter. Respir Care 2002;47(6):682–687.
28. Kresch MJ, Lin WH, Thrall RS. Surfactant replacement therapy. Thorax 1996;51(11):1137–1154.
29. Nelle M, Zilow EP, Linderkamp O. Effects of high-frequency oscillatory ventilation on circulation in neonates with pulmonary interstitial emphysema of RDS. Inten Care Med 1997;23(6): 671–676.
30. Goddard P, et al. Use of continuously recording intravascular oxygen electrode in the newborn. Arch Dis Child 1974;49(11): 853–860.
31. Weiss IK, Fink S, Harrison R, Feldman JD, Brill JE. Clinical use of continuous arterial blood gas monitoring in the pediatric intensive care unit. Pediatrics 1999;103:440–445.
32. Morgan C, et al. Continuous neonatal blood gas monitoring using a multiparameter intra-arterial sensor. Arch Dis Child Fetal Neonatal Ed 80(2):F93–F98.
33. Rais-Bahrami K, Rivera O, Mikesell GT, Short BL. Continuous blood gas monitoring using in-dwelling optode method: Comparison to intermittent arterial blood gas sampling in ECMO patients. J Perinatol 2002;22(6):472–474.
34. Widness JA, et al. Clinical performance of an in-line point-of-care monitor in neonates. Pediatrics 2000;106(3): 497– 504.
35. Gardner RM. Invasive pressure monitoring. In: Civetta JM, Taylor RW, Kirby RR. editors Critical Care, 3rd ed. Philadelphia: Lippincott-Raven; 1997. pp 839–845.
36. http://www.utahmed.com/deltran.htm.
37. Ewert P, Nagdyman N, Fischer T, Gortner L, Lange PE. Continuous monitoring of cardiac output in neonates using an intra-aortic Doppler probe. Cardiol Young 1999;9(1): 42–48.

38. Tsai-Goodman B, et al. Development of a system to record cardiac output continuously in the newborn. Pediatr Res. 1999;46(5):621–625.
39. Hermansen MC, Hermansen MG. Intravascular catheter complications in the neonatal intensive care unit. Clin Perinatol; 2005;32(1):141–156.
40. Barrington KJ. Umbilical artery catheters in the newborn: Effects of position of the catheter tip. Cochrane Database Syst Rev CD000505, 2000.
41. Kempley ST, Bennett S, Loftus BG. Randomized trial of umbilical arterial catheter position: Clinical outcome. Acta Paediatr 1993;82:173–176.
42. Mokrohisky ST, Levine RL, Blumhagen JD. Low positioning of umbilical-artery catheters increases associated complications in newborn infants. N Engl J Med 1978;299:561–564.
43. Umbilical Artery Catheter Trial Study Group. Relationship of intraventricular hemorrhage or death with the level of umbilical artery catheter placement: A multicenter randomized clinical trial. Pediatrics 1992;90:881–887.
44. Haldeman S, Fowler GW, Ashwal S. Acute flaccid neonatal paraplegia: A case report. Neurology 1983;33:93–95.
45. Cumming WA, Burchfield DJ. Accidental catheterization of internal iliac artery branches: A serious complication of umbilical artery catheterization. J Perinatol 1994;14:304–309.
46. Klaus MH, Fanaroff AA. Care of the high-risk neonate, 5th ed. Philadelphia: WB Saunders; 2001.
47. Nowlen TT, Rosenthal GL, Johnson GL. Pericardial effusion and tamponade in infants with central catheters. Pediatrics 2002;110:137–142.
48. O'Grady NP, et al. Guidelines for the prevention of intravascular catheter-related infections. Pediatrics 2002;110(5):e51.
49. Tudos AJ, Besselink GAJ, Schasfoort RBM. Trends in miniaturized total analysis systems for point-of-care testing in clinical chemistry. Lab on a Chip: 2001;1(2):83–95.
50. Wolfbeis OS. Fiber-optic chemical sensors and biosensors. Anal Chem 2001;74(12):2663–2677.
51. Zhang XC. Terahertz wave imaging: Horizons and hurdles. Proceedings of the First International Conference on Biomedical Imaging and Sensing Applications of THz Technology, Physics in Medicine Biology 2002;47(21):3667–3677.

See also ARTERIES, ELASTIC PROPERTIES OF; BLOOD GAS MEASUREMENTS; FIBER OPTICS IN MEDICINE; NEONATAL MONITORING; STRAIN GAGE.

MONOCLONAL ANTIBODIES

BRENDA H. LASTER
JACOB GOPAS
Ben Gurion University of the Negev
Beer Sheva, Israel

INTRODUCTION

This article outlines the association of antibodies within the human immune system, the structural and binding characteristics of antibodies, and the development and production of monoclonal antibodies. Recent advancements in recombinant DNA techniques and genetic engineering are described, including the use of plants to increase the production capacity of Mabs. Their usefulness as biological and medical reagents is further elaborated in a description of the various instrumentation, techniques, and assays employed in the diagnosis and treatment of diseases as well as their utility in the research laboratory.

THE IMMUNE SYSTEM AS IT APPLIES TO ANTIBODIES (1)

The immune system is normally directed at foreign molecules borne by pathogenic microorganisms. However, the immune system can also be induced to respond to simple nonliving molecules. Any substance that can elicit an immune response is said to be immunogenic and is called an immunogen. There is a clear operational distinction between an immunogen and an antigen. An antigen is defined as any substance that can bind to a specific antibody (see below), but is not necessarily able to elicit an immune response by itself.

Immunization

The deliberate induction of an immune response is known as immunization. To determine whether an immune response has occurred and to follow its course, the immunized individual is usually monitored for the appearance of antibodies directed at the specific antigen. Monitoring the antibody response usually involves the analysis of relatively crude preparations of sera. The serum is the fluid phase of clotted blood, which contains a variety of specific antibodies against the immunizing antigen as well as other soluble serum proteins.

Cells Participating in an Immune Response

B lymphocytes (or simply B cells) are one of the two major types of lymphocytes that enable the adaptive immune response. When activated, B cells differentiate into plasma cells that secrete antibodies. T lymphocytes or T cells consist of three main classes. One class differentiates upon activation into cytotoxic T cells, which may kill foreign tissues, cancer cells, and cells infected with virus. The second class of T lymphocytes is T helper cells that differentiate into cells that activate and enable the proper function of other cells, such as B cells. The third class is the T suppressor cells that limit the extent of the immune response.

Antigen Recognition

Both T and B lymphocytes bear receptor proteins on their surface that allow them to recognize antigen. Collectively, these receptors are highly diverse in their antigen specificity, but each individual lymphocyte is equipped with membrane-bound receptors that will recognize only one particular antigen. Each lymphocyte therefore recognizes a different antigen. Together, the receptors of all the different lymphocytes are capable of recognizing a very wide diversity of antigens, which encompass most of the different antigens an individual will meet in a lifetime. These include those antigens that are exclusively synthesized in the laboratory. The B cells do not secrete antibody until they have been stimulated by specific antigen. The B-cell antigen receptor (BCR) is a membrane-bound form of the same antibody that they will secrete when activated by antigen. Thus the antigen recognized by both the BCR and the secreted antibody present in the same B cell, are identical.

Antibodies

Antibody molecules as a class are now generally known as immunoglobulins (Ig), and the antigen receptor of B lymphocytes is known as surface immunoglobulin. The T cell antigen receptor (TCR) is related to immunoglobulins, but is quite distinct from it in structure and function.

Structure and Function of Antibodies

Antibodies are the antigen-specific products (proteins) secreted by B cells. The antibody molecule has two separate functions: one is to bind specifically to molecules from the immunogen (pathogen) that elicited the immune response; the other is to recruit various cells and molecules in order to remove and destroy the pathogen once the antibody is bound to it. These functions are structurally separated in the antibody molecule. One region of the antibody specifically recognizes antigen and the other engages the effector mechanisms that will dispose of it.

The antigen-binding region varies extensively among antibody molecules and is thus known as the variable region or V region, labeled as VH (heavy chain) and VL (light chain). It is this variability that allows each antibody molecule to recognize and bind a particular antigen. The total repertoire of antibodies made by a single individual is large enough to ensure that virtually any structure can be bound. The association between the antibody and the antigen depends on their steric conformation. That is, depending on the size and interatomic distance of these reacting molecules, a tight fit between the antibody combining sites and the antigenic determinant can occur.

The region of the antibody molecule that engages the effector functions of the immune system, but is not associated with antibody binding and does not vary in the same way is known as the constant region or C region. It is typified by the IgG antibody shown in Fig. 1 and is designated CL (light chain) and CH (heavy chain). It has five main forms, or isotypes, that are specialized for activating the different immune effector mechanisms.

The remarkable diversity of antibody molecules is the consequence of a highly specialized mechanism by which

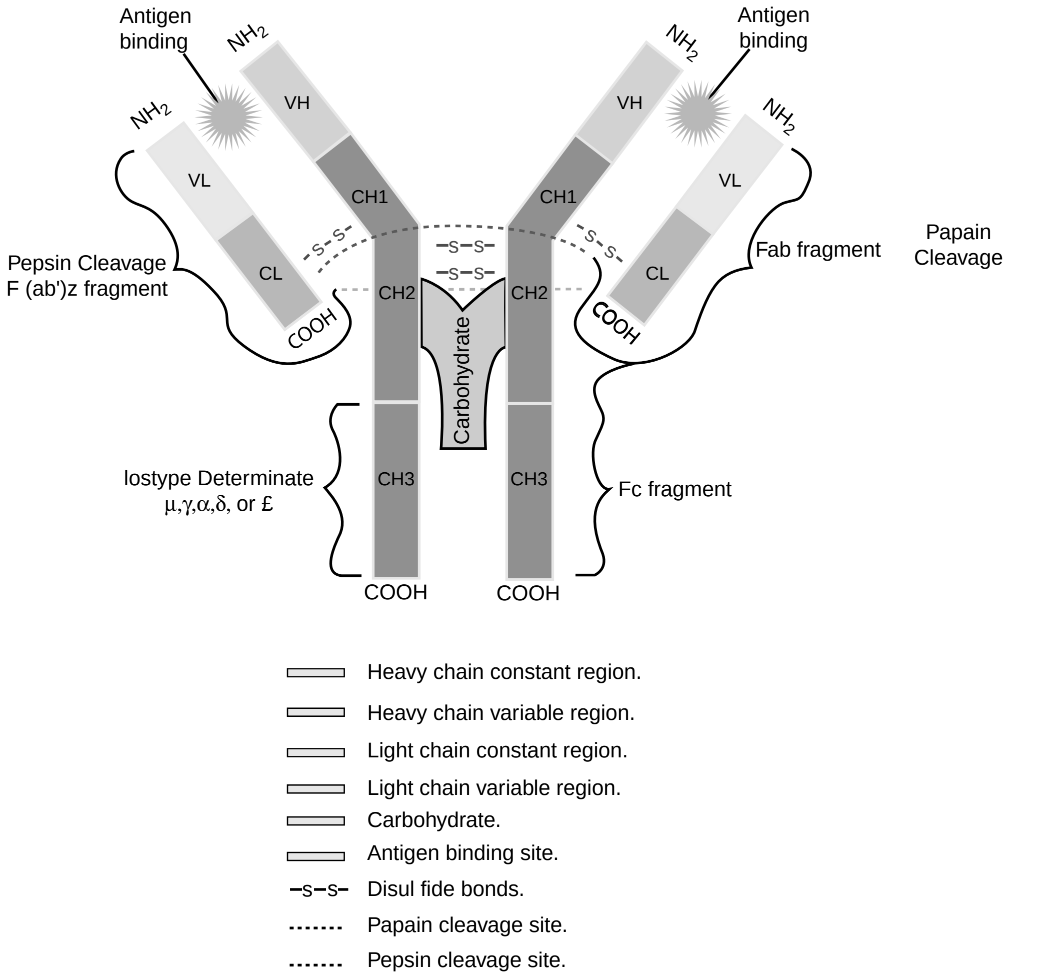


Figure 1. Basic immunoglobulin structure. [Courtesy of Sigma-Aldrich (www.sigmaaldrich.com/img/assets/8181/AntibodyExp).]

the genes that code for antibody production and that are expressed in any given B cell are assembled by DNA rearrangements that join together two or three different segments to form a V region gene during the development of the B cell. Subsequent DNA rearrangement can attach the assembled V-region to any C-region gene and thus produce antibodies of any of the five isotypes.

Antibodies are roughly Y-shaped molecules. All antibodies are constructed in the same way from paired heavy and light polypeptide chains. In Fig. 1, one can observe that the innermost regions of the Y-shaped molecule are the heavy chains; the light chains are the outermost regions. Within this general category, however, five classes (isotypes) of immunoglobulin -IgM, IgD, IgG, IgA, and IgE- can be distinguished biochemically as well as functionally. The five classes are defined by the structure of their heavy chain. Their distinctive functional properties are conferred by differences in the amino acid sequences of the carboxy-terminal part of the heavy chain (COOH in Fig. 1) in the region that is not associated with the light chain. IgG (Fig. 1) will be used to describe the general structural features of immunoglobulin molecules.

The IgG antibodies are large molecules (~150 kDa) composed of two different polypeptide chains. One of these polypeptide chains, ~50 kDa in size, is termed the heavy or H chain. The other, is 25 kDa in size, and is termed the light chain or L chain. The two chains are present in an equimolar ratio, and each IgG molecule contains two heavy chains and two light chains. The two heavy chains are linked to each other by disulfide bonds and each heavy chain is linked to a light chain by a disulfide bond. In any one immunoglobulin molecule, the two heavy chains and the two light chains are identical, enabling them to bind two identical antigenic determinants.

The amino-terminal sequences of both the heavy and light chains vary greatly among different antibodies. The variability in sequence is limited to approximately the first 110 amino acids on the chain, corresponding to the first domain, whereas the carboxy-terminal sequences are constant between immunoglobulin chains, either light or heavy, of the same isotype.

Fragmentation of Antibodies

The antibody molecule can be readily cleaved by different proteases into functionally distinct fragments. For example, Fab fragments, each of which consists of two identical fragments, each containing the antigen binding region. Additionally, an Fc fragment can be extracted that interacts with effector molecules and cells, or one $F(ab')_2$ fragment that contains both arms of the antigen binding region. Figure 1 shows the sites of the derivation of these fragments. Genetic engineering techniques now permit the construction of designed variations of the antibody molecule such as a truncated Fab that comprises only the V region of a heavy chain linked to a V region of a light chain. This is called a single–chain Fv. A broad range of genetically engineered molecules are now becoming valuable therapeutic agents because their smaller size readily permits their penetration into tissue. Useful antibodies from animal sources have been engineered in a process referred to as “humanization”. This avoids their recognition as foreign, and prevents their rapid clearance from the body. The process utilizes the variable region of a mouse antibody coupled to the Fc region from human antibodies. Antibody fragments may also be coupled to toxins, radioactive isotopes and protein domains that interact with effector molecules or cells.

MONOCLONAL ANTIBODIES

Antibody Heterogeneity

The antibodies generated in a natural immune response or after immunization in the laboratory are a mixture of molecules of different antigen specificities and affinities. Because of their multiple specificities they are termed polyclonal antibodies. Some of this heterogeneity results from the production of antibodies that bind numerous different antigenic determinants (epitopes) present on the immunizing antigen. However, even antibodies directed at a single antigenic determinant can be markedly heterogeneous. Antisera (serum containing antibodies against specified antigens) are valuable for many biological purposes, but they have certain inherent disadvantages that relate to the heterogeneity of the antibodies they contain. First, each antiserum is different from all other antisera, even when raised in a genetically identical animal while using the identical preparation of antigen and immunization protocol. Second, antisera can be produced in only limited volumes, and thus it is impossible to use the identical serological reagent in a long or complex series of experiments, or in clinical tests or therapy. Finally, even purified antibodies may include minor populations of antibodies that give unexpected cross-reactions that confound the analysis of experiments and can be harmful in therapy. To avoid these problems, and to harness the full potential of antibodies, it became necessary to develop a method for making an unlimited supply of antibody molecules of homogeneous structure and known specificity. This has been achieved through the production of monoclonal antibodies from hybrid antibody forming cells or, more recently, by genetic engineering.

Production of Monoclonal Antibodies

In 1975, using cell culture techniques, Kohler and Milstein (2) found a way to grow an immortal B-cell-like lymphocyte that continuously produced an antibody with a predetermined specificity. The procedure required the immunization of a mouse in order to produce a large population of B-cells in the spleen that would secrete a specific antibody. However, in cell culture, the life span of spleen cells is only a few days. To produce a continuous source of antibody, the B cells had to grow continuously. This was achieved by fusing the spleen cells that contained a particular gene, the hypoxanthine-guanine phosphoribosyl transferase (HGPRT) gene with immortal myeloma cells (cancerous plasma cells). In general, plasma cells are mature-antibody secreting B cells. The myeloma cells were preselected to ensure three specific properties. First, immortality in cell culture; second, that they were sufficiently altered so that

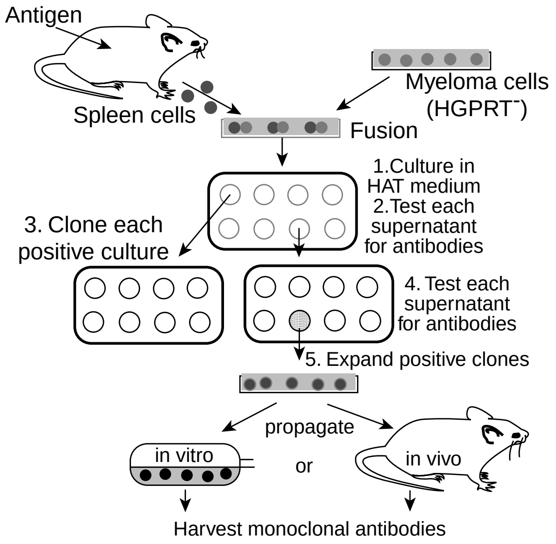

Antigen
Spleen cells
Myeloma cells (HGPRT⁻)
Fusion
1.Culture in HAT medium
2.Test each supernatant for antibodies
3. Clone each positive culture
4. Test each supernatant for antibodies
5. Expand positive clones
propagate
in vitro
or
in vivo
Harvest monoclonal antibodies

Figure 2. Schematic of hybridoma protocol. [Courtesy of Prof. John Kimball (http://users.rcn.com/jkimball.ma.ultranet/Biology Pages/M/Monoclonals.html).]

they did not secrete antibody, and third, that they would not flourish in a particular growth medium containing hypoxanthine, aminopterin and thymidine (HAT). The HAT medium was previously shown to be a highly selective medium for those specific hybrid cells that lack the gene for the enzyme, HGPRT. Consequently, all unfused myeloma and spleen cells would not survive in the HAT medium. The HGPRT gene that was contributed by the spleen cell permitted only hybrid cells to survive in the HAT medium, because only hybrid cells would be able to grow in the culture due to the conferral of immortality by the myeloma cells. Therefore, the immune spleen cells conferred both antibody specificity and the HGPRT gene to the hybrid cell, while the myeloma cell conferred immortality to the spleen cell and they were able to survive indefinitely in culture. This is the method that is used today to produce individual hybridomas (the hybrid cells) that are then screened for antibody production. Single antibody-producing cells that produce an antibody with the desired specificity are cloned. These cloned hybridoma cells are grown in bulk culture to produce large amounts of antibody that are used in a variety of ways. Since each hybridoma is descended from a single B cell, all cells of a particular hybridoma cell line produce the same hybridoma molecule. This is the monoclonal antibody (Mab). The procedure is shown as a schematic outline in Fig. 2.

Recombinant Genetic Engineering

The field of antibody engineering endeavors to improve the target specificity and effector function of the antibodies by altering their construction, while retaining their binding characteristics. This is achieved by using new recombinant engineering technologies to redesign the molecular architecture of the antibodies. Examples of this are shown in Fig. 3, where in a process called genetic engineering, recombinant DNA (spliced DNA that is formed from two or more sources and then rejoined) is produced by putting genes from one species into the cells of an individual organism of another species. The foreign DNA becomes

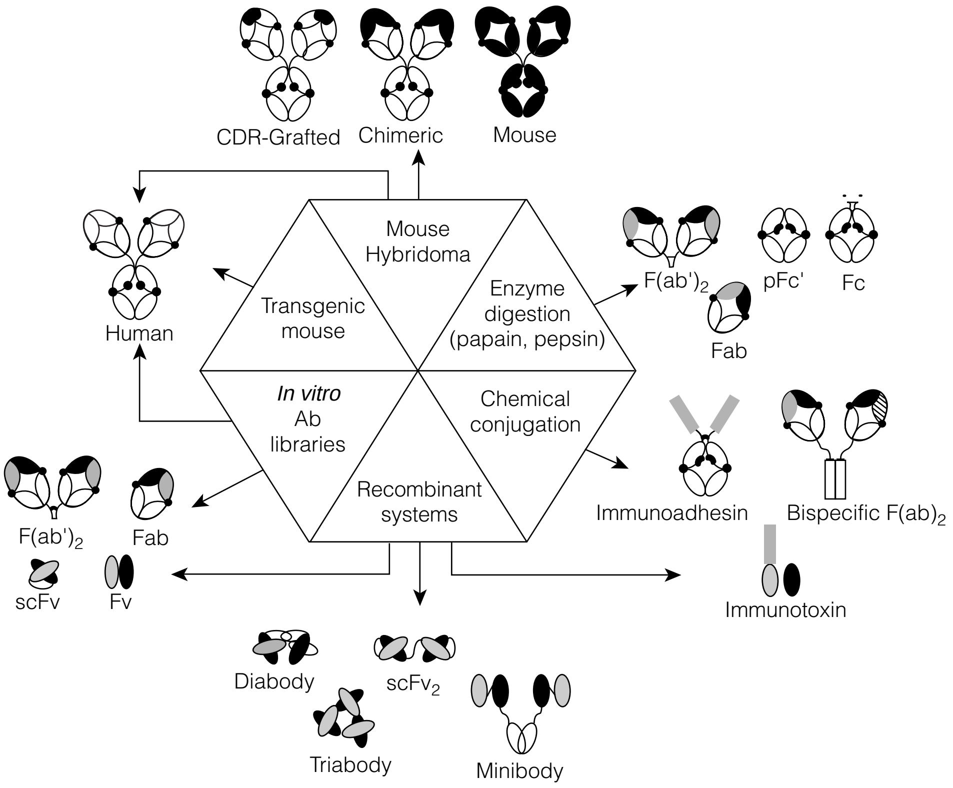


Figure 3. Antibody engineering: examples of antibody-based constructs and respective routes for their generation. Hybridoma technology provides a way of producing mouse monoclonal antibodies. Genetic engineering has encouraged the creation of chimeric, humanized, and human antibodies, antibody fragments (traditionally obtained by partial digestion of immunoglobulins with proteases), multimeric antibody fragments, fusion (immunoadhesins and immunotoxins) and bispecific antibodies. Multimeric antibody fragments (diabody, triabody, and tetrabody) are represented as multivalent structures, although they can also be engineered to be multispecific. The minibody depicted is a dimer that can be linked to the CH_3 fragment via a LD linker or a flexible linker (FlexMinibody). Bispecific F(ab)2 is shown as a Fab dimer linked noncovalently via interaction of amphipathic helices. (Courtesy of American Chemical Society and American Institute of Chemical Engineers, Copyright © 2004.) (See Ref. 3.)

part of the host's genome (its genetic content) and is replicated in subsequent generations descended from that host. An alternative technique for producing antibody-like molecules is the Phage Display Libraries for Antibody V-region Production (4). In this approach, gene segments that encode the antigen-binding variable region of antibodies are fused to genes that encode the coat protein (outside surface) of a bacteriophage (viruses that infect bacteria). In essence, mRNA from primed human B-cells is converted to cDNA. The large variety of diverse antibody genes are expanded by the polymerase chain reaction (PCR) to generate a highly diverse library of antibody genes. Bacteriophage containing such gene fusions are then used to infect bacteria, resulting in phage particles that have outer surfaces that express the both antibody-like fusion protein, and the same antigen-binding domain displayed on the outside of the bacteriophage. The collection of such recombinant phages, each displaying a different antigen-binding domain on its surface, is known as a phage display library. A particular phage can be isolated from the mixture and can be used to infect fresh bacteria. Each phage isolated in this way produces a monoclonal antigen-binding particle analogous to a monoclonal antibody. A complete antibody molecule can then be produced by fusing the V region of a particular phage with the invariant part of the immunoglobulin gene. These reconstructed antibody genes can be introduced (transfected) into a suitable host cell line. These genes will become part of the cell's genome and will secrete antibodies akin to hybridomas.

Monoclonal Antibodies Produced in Plants, Plantibodies

After undergoing genetic engineering techniques, plant cells are capable of assembling and producing unlimited quantities of antibodies, referred to as plantibodies (5). This trademark name for human antibodies manufactured in plants has functionally limitless production capacity and lower costs than those associated with the yeast fermentation process that is currently being used to produce mass quantities of human antibodies. This fairly recent finding might prove to be of benefit in the medical, consumer, and industrial applications of monoclonals. For example, it has been postulated that the development of plantibodies with a capability of sequestering heavy metals or radioactive compounds might have a very positive impact on the environment, particularly because their production is very inexpensive and large supplies are easily produced. Because the corn crop is so readily available worldwide, and its kernel stores natural plantibodies, these can be purified as needed by standard milling procedures. Potato and tomato crops are also being used. The first clinical use of the effectiveness of a plantibody was against the bacterium, *Streptococcus mutans*. This organism produces lactic acid that erodes tooth enamel. The plantibody was brushed onto human teeth for 3 weeks and tooth decay was prevented for up to 4 months. The action of the antibody was to prevent the bacterium from binding to the tooth surface. Plantibody-containing gels are being developed to prevent genital herpes infections and to protect newborn babies during delivery against transmission of the Herpes virus from infected mothers. Plantibodies against human immunodeficiency verus (HIV) and the production of sperm are also being developed. Concerns have been expressed about the use of genetically engineered food crops because of the potential dangers of their getting into the wrong hands, or disturbing the ecological balance.

The aforementioned technologies have revolutionized the use of antibodies by providing a limitless supply of antibodies with single and known specificity. Monoclonal antibodies are now used in most serological assays, as diagnostic probes and as therapeutic agents.

TECHNIQUES FOR USING MONOCLONAL ANTIBODIES AS SEROLOGICAL AND DIAGNOSTIC PROBES

Monoclonal antibodies can serve as tools for diagnosing and treating disease and are valuable agents in the research laboratory. Their utility required the development of procedures that would permit them to be viewed at the particular region of interest. Some of the most widely used techniques are described in the following sections (1).

Immunofluorescence

Since antibodies bind stably and specifically to their corresponding antigen, they are invaluable as probes for identifying a particular molecule in cells, tissues, or biological fluids. Monoclonal antibody molecules (Mabs) can be used to locate their target molecules accurately in single cells or tissue sections by a variety of different labeling techniques. When either the antibody itself, or the anti-Mab that is used to detect it, is labeled with a fluorescent dye the technique is known as immunofluorescence. As in all serological techniques, the antibody binds stably to its antigen, allowing any unbound antibody to be removed by thorough washing. The fluorescent dye can be covalently attached directly to the specific antibody, but more commonly, the bound antibody is detected by a secondary fluorescent anti-immunoglobulin; that is, the first antibody binds to the antigen and a fluorescent secondary antibody (antibody) is targeted to the primary antibody–antigen complex. The technique is known as indirect immunofluorescence, which is demonstrated in Fig. 4, where the binding of the first antibody to the antigen is followed by the binding of the antibody. The dyes chosen for immunofluorescence are excited by light of a particular wavelength, and emit light of a different wavelength in the visible spectrum. By using selective filters that can permit only certain wavelengths of light to pass, only that light coming from the dye or fluorochrome used is detected in the fluorescence microscope. Therefore, the antibody can be located by virtue of its emission of fluorescent light. The recently developed confocal fluorescent microscope considerably enhances the resolution of the technique. If different dyes are attached to different antibodies, the distribution of two or more Mabs can be determined in the same cell or tissue section. Differentiating between the antibodies occurs because either the dye or the fluorochrome will excite at different wavelengths or because they will emit their fluorescence at different wavelengths. An example of the immunofluorescence technique is shown in Fig. 4, whereby, through the use of monoclonal antibodies targeted to

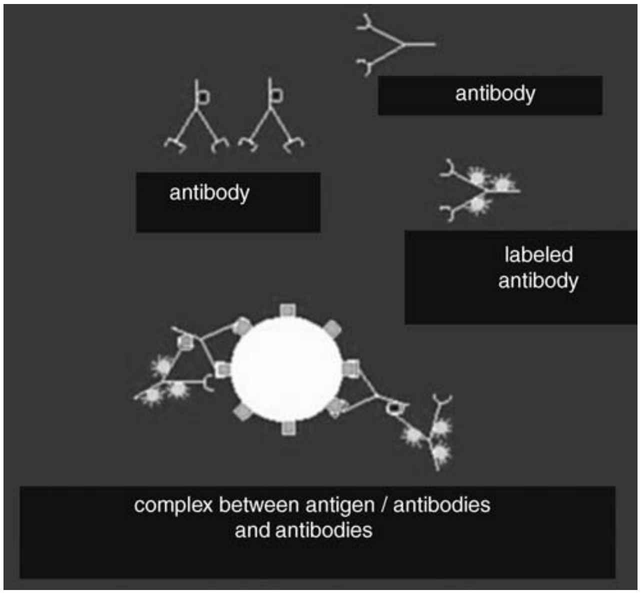


Figure 4. Indirect immunofluorescence. The primary antibody binds to the antigen and the fluorescent antiantibody binds to the primary thereby increasing the signal. [Courtesy of Prof. v. Sengbusch (http://www.biologie.uni-hamburg.de/b-online/d00/copyrig.htm).]

intracellular proteins, certain structures within the cell become visible. The structure shown in Fig. 5 is that of the spindle, which appears during cell division. During certain phases of cell division, the chromosomes arrange themselves in the equatorial plane of the spindle. The spindle is made up of microtubules that, in turn, are composed of proteins. Monoclonal antibodies that would bind to two specific proteins, α and γ-tubulin, of the microtubule were synthesized and labeled with two different fluorochromes. During cell division, the cells were exposed to the fluorescent-labeled Mabs that formed a complex with the proteins and permitted visualization of the spindle.

Immunohistochemistry

An alternative to immunofluorescence for detecting a protein in tissue sections is immunohistochemistry, in which the specific Mab is chemically coupled to an enzyme that converts a colorless substrate into a colored reaction product *in situ*. The Enzyme-Linked ImmunoSorbent Assay (ELISA) is a technique that detects and quantifies specific antigens from a mixture. It is widely used in procedures that screen blood for viral or bacterial contamination, to detect infections, toxins, illegal drugs, or allergens, and in measuring hormone levels, such as in pregnancy or thyroid function. The assay involves the binding of an antibody to a solid surface and exposing it to the antigens. A second complex, consisting of the same antibody, but additionally tagged with a particular enzyme, is exposed to the initial antibody–antigen conjugate and binds. After washing the surface to remove excess unbound antigen, a colorless substrate is added that permits the antigen to be converted into a colored product that can be read and measured using absorption spectrometry. The intensity of color is proportional to the concentration of bound antigen. A schematic of the ELISA assay is shown in Fig. 6. A more detailed description of the Elisa procedure can be found in (Kimball's Biology Pages http://biology-pages.info). Horseradish peroxidase and alkaline phosphatase are also used as enzymes in immunochemistry assays. An example of the utility of these enzymes for protein detection is shown in Fig. 7.

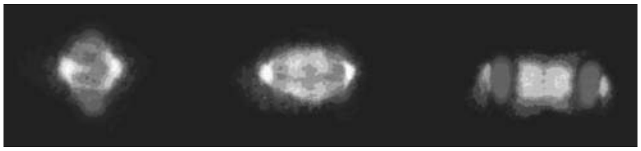


Figure 5. Dividing cells (mitosis: metaphase, anaphase, and telophase) were stained with monoclonal antibodies against two intracellular proteins, α-tubulin in green, and γ-tubulin in red. Because these proteins constitute the spindle, the intracellular structure upon which chromosomes line up during mitosis, the structure is visualized by virtue of the difference in the fluorochromes tagged to the Mabs that were bound to the proteins. Chromosomes were stained with a blue dye. (www.img.cas.cz/dbc/gallery.htm.)

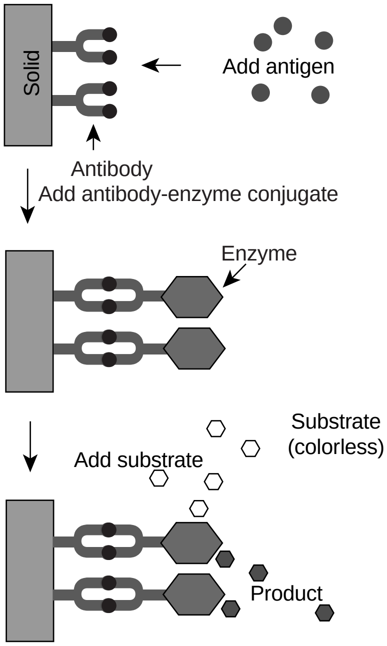


Figure 6. Schematic of ELISA assay protocol. [Courtesy of Prof. John Kimball (http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/E/Elisa.html).]

Immunoelectron Microscopy

Antibodies can be used to detect the intracellular location of structures or particular molecules by electron microscopy, a technique known as immunoelectron microscopy. After labeling Mabs with gold particles and targeting them to samples, they can then be examined in the transmission electron microscope. Since electrons do not penetrate through gold particles, the regions in which the antibodies bind appear as dark dots.

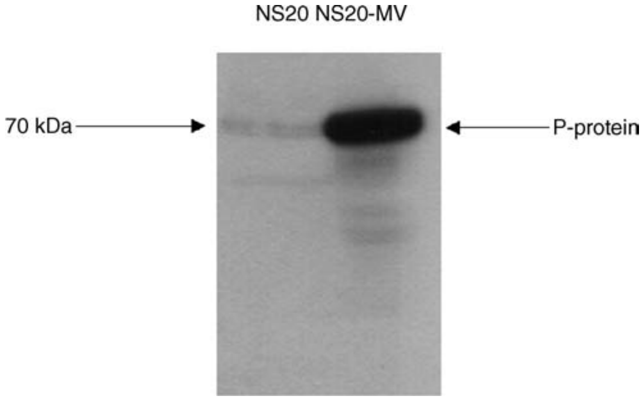


Figure 7. Detection of the measles virus (MV) P-protein by Western Blot in MV infected and noninfected cells. Whole-cell lysates were prepared from MV that was either persistently infected (NS20-MV) or notinfected (NS20) mouse neuroblastoma cells. The proteins in the lysates were separated by SDS–PAGE and blotted onto nitrocellulose paper. The blot was incubated with a Mab against the MV P-protein, followed by a secondary antimouse immunoglobulin antibody linked to horseradish peroxidase. The P-protein band was detected when a substrate was added that was modified by peroxidase on the blot and caused light to be released. Light was detected on a specific band after exposure to film. The results show that only the measles- infected cells express the viral protein. (Courtesy of Jacob Gopas.)

Blotting Techniques

Immunoblotting or Western blotting is used to identify the presence of a given protein in a cell lysate. Cells are placed in detergent to solubilize all cell proteins and the lysate (the material resulting from the ruptured cells) is run on sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS–PAGE), which enables protein migration and separation by size. Further resolution is achieved if the proteins are initially separated by charge according to their isoelectric point and then by size. This technique is referred to as two-dimensional (2D) gel electrophoresis. The proteins are then transferred (blotted) from the gel to a stable support, such as a nitrocellulose membrane for easier handling. Specific proteins of interest in the lysate's mixture are detected by incubating the membrane with a Mab that can react with a defined protein on the membrane. An example of the technique is shown in Fig. 7. The proteins bound to the antibodies are revealed by enzyme-labeled, anti-immunoglobulin antibodies. By this technique the presence or absence, as well as the amounts of specific proteins, can be monitored following a variety cell treatments. Specific DNA labeled with antigen (hapten)-bound nucleotides can be blotted onto a membrane and detected with Mabs against the hapten. This allows the detection of viral or bacterial DNA in tissues or body fluids, as generated by PCR.

Purification Techniques

Affinity chromatography and immunoprecipitation are techniques that enable purification of molecules and their characterization. A mixture of molecules can be incubated with a Mab, which is chemically attached to a solid support. The bound antibody–antigen complex is washed from unbound molecules by centrifugation, and then the molecule of interest is eluted for further characterization. These techniques are useful for protein purification, for determining its molecular weight, its abundance, distribution, and whether it undergoes chemical modifications as a result of processing within the cell.

Immunoelectrophoresis

Two-dimensional electrophoresis is used to separate different antigens that might be present in one solution. The antigens are separated on the basis of their electrophoretic mobility. The currents are run at right angles to each other, driving the antigens into the antiserum (containing Mabs). Peaks are obtained when the antigen forms a complex with the antibody; the area under the peaks gives the concentration of antigen as shown in Fig. 8. Rocket electrophoresis is a similar technique. Here, after a current is applied, the antigens are separated based upon their ionic charge by their differential migration through a gel that contains antibody. As shown in Fig. 9, concentration is determined by the migration distance. In countercurrent electrophoresis, the greater internal osmotic pressure drives the antibody backwards into a gel after a current is applied. An antigen that is negatively charged will form a complex with the antibody in the gel in a pH-dependent process.

Instrumentation

An immensely powerful tool for defining and enumerating and isolating cells is the use of the fluorescence-activated

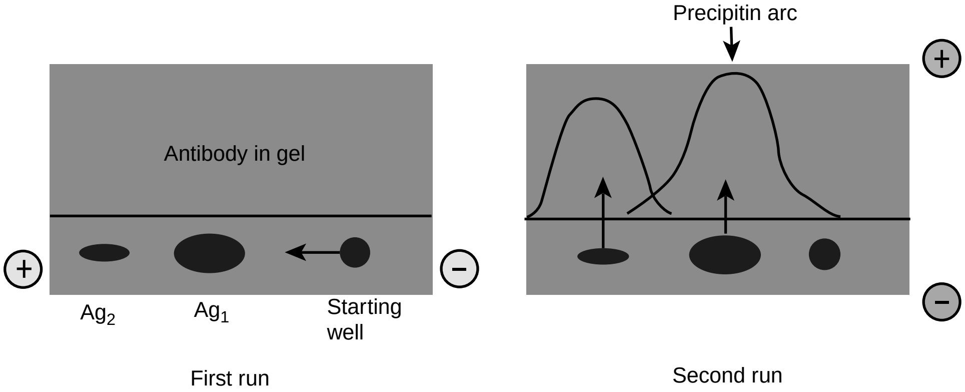


Figure 8. Two-dimensional immunoelectrophoresis. Antigens are separated on the basis of electrophoretic mobility. [Courtesy of the Natural Toxins Research Center at Texas A&M University – Kingsville (http://ntri.tamuk.edu/).]

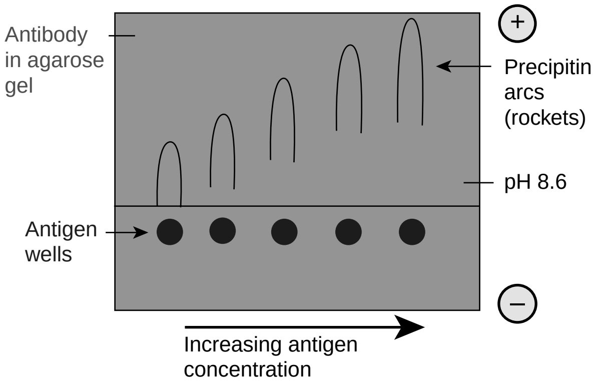


Figure 9. Rocket electrophoresis. Antigen is electrophoresed into gel containing antibody. The distance from the starting well to the front of the rocket shaped arc is related to antigen concentration. [Courtesy of the Natural Toxins Research Center at Texas A&M University – Kingsville (http://ntri.tamuk.edu/).]

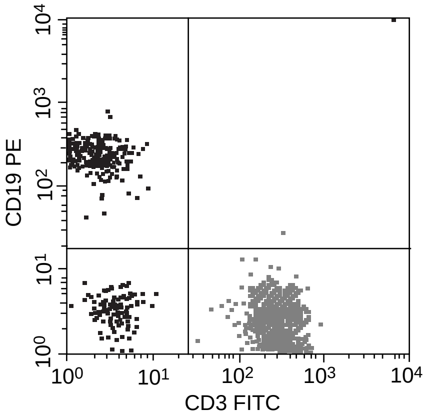


Figure 10. FACS analysis. Characterizing cells at different stages of development through the use of fluorescent labeled monoclonal antibodies against cell surface markers is one of the most common applications of flow cytometry. Changes in the relative numbers, absolute counts, or in the ratio of cell types can provide valuable information as to the status of the immune system in human disorders or animal models. Different cell types can be detected and quantified from a mixed population by the use of monoclonal antibodies labeled with different fluorescent dyes that have nonoverlapping emission spectra. In this example experiment, blood lymphocytes were incubated with two different Mabs, CD19, and CD3. CD19 was labeled with the fluorochrome phycoerythrin (PE) and binds a cell membrane molecule specific for B-lymphocytes. The CD3 was labeled with fluorecein isothyocyanate (FITC) that detects a cell membrane protein specific for T-lymphocytes. Three populations of cells were detected in this experiment according to the antibody bound to the cells. The logarithmic *x* and *y* axis represent relative amounts of fluorescence detected on cells labeled with FITC or PE, respectively. The blue dots represent cells unstained by either of the antibodies, the red dots represent B-lymphocytes that were detected by CD19 and the green dots represent T-lymphocytes, CD3 positive cells. No cells were detected that bound both antibodies (top right quadrant).

cell sorter (FACS). This instrument is used to study the properties of cell subsets identified using Mabs to cell surface proteins. Individual cells are first tagged by treatment with specific fluorescent Mabs. The mixture of labeled cells is then forced with a much larger volume of fluid through a nozzle, creating a fine stream of liquid containing cells spaced singly at intervals. As each cell passes through a laser beam it scatters the laser light, and any dye molecules bound to the cell will be excited and fluoresce. Sensitive photomultiplier tubes detect both the scattered light, which gives information on the size and granularity of the cell, and the fluorescence emission, provide quantification of the binding of the labeled Mabs, and on the expression of cell-surface proteins by each cell. In the cell sorter, the signals passed back to the computer are used to generate an electric charge, which is passed from the nozzle through the liquid stream. Droplets containing a charge can then be deflected from the main stream as they pass between plates of opposite charge. In this way a specific population of cells, distinguished by the binding of the labeled antibody and its defined electrical charge, can be extracted and purified from a mixed population of cells. Alternatively, to deplete a population of cells, a labeled antibody directed at marker proteins expressed by undesired cells will direct the cells to a waste channel, retaining only the unlabeled cells. Several Mabs labeled with different fluorochromes can be used simultaneously. FACS analysis can give quantitative data on the percentage of cells bearing different molecules, and the relative abundance of the particular molecules in the cell population, 10,000 cells in a typical experiment demonstrates the retrieval of data after FACS analysis. An example of data output from FACS is shown in Fig. 10.

Mabs as Molecular Probes

Mabs can also be used to determine the function of molecules. Some antibodies are able to act as agonists, when the binding of the Mab to the molecule mimics the binding of the natural ligand (antigen) and activates its function. For example, antibodies to the CD3 antigen present on mature human T cells have been used to stimulate the T cells. This occurs because CD3 is associated with the T-cell receptor and is responsible for signal transduction of the receptor. Conversely, Mabs can function as antagonists, inhibiting the binding of the natural ligand and thus blocking its function. For example, antibodies that block the epidermal growth factor receptor (a growth stimulating protein) function as antagonists.

THE USE OF MONOCLONAL ANTIBODIES AS THERAPEUTIC AGENTS

Mabs against cell-surface molecules have been used to remove specific lymphocyte subsets or to inhibit cell function *in vitro*. Cytotoxic drugs kill proliferating cells indiscriminately. In contrast, antibodies can interfere with immune responses in a nontoxic and much more specific manner. For example, Mabs can be used to remove undesirable lymphocytes from donor bone marrow cells prior to transplantation. This treatment selectively removes lymphocytes that recognize the host tissues as “foreign” and induce a potentially fatal condition known as Graft versus Host reaction (6).

Mabs are being tested experimentally to inhibit transplant rejection, to alleviate and suppress autoimmune disease and in cancer detection and treatment. The major impediment to therapy with monoclonal antibodies in humans is that these antibodies are mostly of mouse origin, and humans rapidly develop antibody responses to mouse antibodies. This not only blocks the actions of the mouse antibodies, but leads to allergic reactions. If this occurs, future treatment of the same patient with any mouse Mab is unacceptable. In principle, the problem can be avoided by producing antibodies that are not recognized as foreign by the human immune system. Several strategies are being explored for their construction. One approach is to clone human V regions into a phage display library (see above) and select for its ability to bind human cells. With this method, Mabs that are entirely human in origin can be obtained. Second, mice that lack endogenous immunoglobulin genes can be made "transgenic" (chimeric). That is, they can have human genes put into their genome through recombinant DNA techniques. When this occurs, they will then express human immunoglobulin heavy and light genes and eventually antibody molecules. A third approach is to graft the variable region of a mouse Mab into the rest of the human immunoglobulin molecule, in a process known as humanization. These recombinant antibodies are far less immunogenic in humans than the Mabs of the parent mouse, therefore, they can be used for more efficient and repeated treatment of humans with far less risks. In some cases even humanized antibodies may evoke an immune response and must be administered with immunosuppressive drugs.

THE USE OF MONOCLONAL ANTIBODIES IN THE DETECTION, FOLLOW-UP, AND TREATMENT OF CANCER

Tumor-Specific Antigens

For the greater part of the twentieth century, it was assumed that any antigens present on the cell surface of tumor cells would also be present in normal cells; therefore, few investigations were undertaken to elicit any autoimmune response against cancer cells. However, once inbred mouse strains bearing transplanted syngeneic (genetically identical) tumors became available, research studies validated that immune reactions against these tumors could be induced with no toxic effects on normal tissues, and scientists began to pursue the identification of "tumor specific antigens". Shared tumor antigens were found in many of the same types of cancers in different patients, and unique antigens were isolated that were specific for a particular cancer in a particular patient. The SEREX database lists the antigens that have been isolated from humans (7). These antigens have the ability to generate an immune response when introduced into a patient.

The advent of monoclonal antibodies suggested the possibility of targeting and destroying tumors by making antibodies against tumor-specific antigens. However, this relies upon the identification of a tumor-specific antigen that is located on the surface of cancer cells. Because of their ability to differentiate between normal and malignant tissues and to exact a variety of antitumor responses, Mabs offer a significant advantage to conventional forms of therapy. Several monoclonal antibodies have already been proven to be relatively well tolerated and effective for the treatment of many different malignant diseases.

Approaches to Cancer Immunotherapy

Approaches to cancer immunotherapy can be either active or passive. For example, in the active category, tumor vaccines that immunize against particularly defined tumor antigens, can be used. In the passive category is the use of monoclonal antibodies that are either conjugated, unconjugated, or radiolabeled. These same approaches can also be categorized as specific, wherein antigens are directly targeted, or nonspecific, where immune cells are used to directly target tumor cells. Other approaches are taken that elicit antitumor effects with different mechanisms, such as using antibodies to block growth factors or receptors on cells; targeting specific tissue components of the tumor or its blood vessels; interfering with cell signals; or with apoptosis (programmed cell death) (8).

Magic Bullets

While such Mab-based therapies offer a high potential to fulfill the promise of "magic bullets" for the treatment of malignant disease, successful application of these therapies is often impaired by several impediments. Factors inhibiting the therapeutic benefit of Mabs may include low or heterogeneous expression of target antigens by tumor cells, high background expression of antigen on normal cells, host antibody immune responses to the Mabs themselves, insufficient anti-tumor response after Mab binding, as well as physical obstructions preventing antibody binding, such as crossing to and from blood vessels as well as tissue barriers en route to the solid tumor mass (9). These factors influence the ability of the Mabs to penetrate to the tumor.

IMAGING TUMORS WITH MONOCLONAL ANTIBODIES

Mabs in Nuclear Medicine

The presence of malignant tumors can be detected through the use of monoclonal antibodies radiolabeled most frequently with the isotopes technetium-99m (^{99m}Tc) or indium-111 (^{111}In). The particular label selected depends upon the size of the antibody. For example, large fragments or whole antibodies require a longer half-life isotope, such as ^{111}In ($T_{1/2} = 2.8$ days), whereas smaller Fab fragments, that are cleared from the body more quickly, can be labeled with ^{99m}Tc ($T_{1/2} = 6$ h). Imaging is performed by a Single Photon Emission Computed Tomography (SPECT) camera whose detectors scan the body and register the radioactive counts. The counts are then mathematically transformed into an image that displays the sites of radioactivity. The nuclear medicine procedure that utilizes this procedure is known as Tumor-Specific Monoclonal Antibody Radioscintigraphy. Because of occasional difficulties with these techniques, such as inadequate tumor perfusion, inadequate amounts of antigen on the surface of the tumor cells, antigen heterogeneity, and nonspecific uptake, new approaches are being investigated. However, due to limited clinical experience, it is too early to predict whether they will improve imaging performance (10). Among these methods is the use of other imaging techniques, such as bone scans or computed tomography (CT), in conjunction with SPECT. In other approaches, attempts are being made to augment surface tumor cell antigens by prestimulation with growth factors, such as cytokines (11).

TREATMENT OF HEMATOLOGICAL MALIGNANCIES

Blood-Cell Cancers

Surface antigens on B- and T-cell lymphocytes are also useful targets for the treatment of blood cell (hematopoietic) malignancies, such as leukemias and lymphomas. These antigens are also expressed at high levels on the surface of various populations of malignant cells, but not on normal tissues. With few barriers present to impede Mab binding, hematologic malignancies are well suited to Mab-based therapy. In recent years, several promising Mab-based therapies for the treatment of hematologic malignancies have been developed and either have already received U.S. Food and Drug Administration (FDA) approval or are in the advanced phases of clinical testing (12). The chimeric antibody, rituxan (rituximAb, Genentech, San Francisco, CA) was among the first Mabs awarded Food and Drug Administration approval for the treatment of non-Hodgkin's lymphoma (13,14). This chimeric (human–mouse) antibody binds CD20, a cell surface antigen expressed on mature B lymphocytes and over 90% of non-Hodgkin's lymphoma cells, but not on hematopoetic progenitor or stem cells. Rituxan has proven to be well tolerated and effective in the treatment of non-Hodgkin's lymphoma either by itself, or in combination with traditional chemotherapy, particularly in patients who are refractory to other types of therapy (15). Campath-1 (alemtuzumAb, Ilex Oncology, San Antonio, TX) is another antibody that has also received FDA approval for the treatment of patients suffering from chronic lymphocytic leukemia. A third Mab to receive FDA approval for the treatment of hematologic malignancies is the chimeric Mab, mylotarg (gemtuzumAb ozogamicin, Wyeth-Ayerst Laboratories, Philadelphia, PA). This antibody targets the CD33 antigen expressed on myeloid (white cells) precursors and leukemic cells, and is absent from normal tissues and pluripotent hematopoetic (blood-cell producing) stem cells.

TREATMENT OF SOLID TUMORS

In comparison to the management of hematologic malignancies, successful treatment of solid tumors with Mabs has proven more elusive; however, some significant therapeutic benefits have been achieved. Herceptin (trastuzumAb, Genentech) is a humanized antibody that has received FDA approval for the treatment of metastatic breast cancer. This Mab recognizes an extracellular domain of the HER-2 protein. Clinical trials with herceptin have shown it to be well tolerated both as a single agent for second or third line therapy, or in combination with chemotherapeutic agents as a first line of therapy. Combination therapy resulted in a 25% improvement of overall survival in patients with tumors that overexpress HER-2, and that are refractory to other forms of treatment (16).

The antiepithelial cellular adhesion tumors Mab molecule, Panorex (eclrecolomAb, GlaxoSmith-Kline, United Kingdom), is another Mab -based therapy that is currently being used for the treatment of colorectal cancer. Panorex has shown tangible benefit for cancer patients and has received approval in Germany for the treatment of advanced colorectal cancer. Like other Mabs used for the treatment of solid tumors, Panorex has proven more efficacious in the treatment of micrometastatic lesions and minimal residual disease in comparison to bulky tumor masses (17).

The failure of Mabs in the treatment of bulky lesions is primarily attributable to the low level of injected Mabs that actually reaches its target within a sizable solid-tumor mass. Studies using radiolabeled Mabs suggested that only a very small percentage of the original injected antibody dose, ~0.01–0.1/g of tumor tissue, will ever reach target antigens within a solid tumor (18). This low level of binding is due to the series of barriers confronted by an administered Mab en route to antigens expressed on the surface of tumor cells.

ELICITING ANTITUMOR RESPONSES

After successfully negotiating the gauntlet of obstacles obstructing access to the target cells within a tumor, a therapeutic Mab must still be capable of eliciting a potent antitumor response. Although it is often ambiguous as to the exact mechanisms by which a particular Mab may mediate an antitumor response, both direct and indirect mechanisms can potentially be involved.

Antibodies of the IgG_1 and IgG_3 isotypes can support effector functions of both antibody-dependent cell-mediated cytotoxicity and complement-dependent cytotoxicity. Antibody-dependent cell-mediated cytotoxicity is triggered by interaction between the Fc region of a cell-bound antibody and Fc receptors on immune effector cells such as neutrophils, macrophages, and natural killer cells. This mechanism is critical for the antitumor effects of several therapeutic Mabs.

Many early studies showed that murine Mabs had limited potential to elicit a potent antitumor response, because the murine Fc regions are less efficient at recruiting human effector cells than their human counterparts. This problem has been largely alleviated by the use of chimeric and humanized antibodies. Genetic engineering techniques have also been used to improve the immunologic effects of therapeutic Mabs by altering antibody shape and size, increasing the valency (bonds of affinity) of Mabs, and creating bifunctional antibodies with two antigenic receptors, one to a tumor antigen and another to an effector cell to increase efficiency of antibody-dependent cell-mediated cytotoxicity (19).

In addition to immunologic effects, Mabs can induce antitumor effects by a variety of direct mechanisms, including the induction of apoptosis (programmed cell death) (20), or the prevention of soluble growth factors from binding their cognate receptors, such as epidermal growth factor (EGF-R) (21) and HER-2 (22). Additionally, Mabs can also be engineered to deliver a cytotoxic agent directly to the tumor. This offers the potential to combine the biological effects of Mabs with the additional effect of a targeted cytotoxic response. The anti-CD33 Mabs, mylotarg, is one such antibody. Combined with the cytotoxic agent, calichaemicin, mylotarg has been reported to be relatively well tolerated, and effective in the treatment of chronic lymphocytic leukemia (23). Antibodies can also be engineered to deliver ionizing radiation directly to tumor cells. Mabs have been conjugated to both α- and β-particle emitting radionuclides (24). Clinical trials in humans also portend the promise of radiolabeled Mabs for the treatment of cancer. In a recent phase III randomized study, patients with relapsed or refractory non-Hodgkin's lymphoma were treated with yttrium-90 and iodine-131 labeled Mabs targeting the CD20 antigen (ibrituximomAb, tiuxetan, and tositumomAb, respectively). Patients treated with these radiolabeled Mabs showed a statistically significant increase in overall response compared with those treated with an unlabeled version of the Mab (rituximAb) (25).

OTHER USES FOR MONOCLONAL ANTIBODIES

Proteomics

After having sequenced the entire human genome, the current task is to understand the "proteome" by identification and quantification of all proteins in a given sample. So far, DNA microarrays have been employed to detect the transcription level [production of messenger ribonucleic acid (mRNA)] of genes in cells. However, it has been found that there is no stringent correlation between transcription level and protein abundance. Furthermore, the status of a protein in terms of its modification and structure cannot be determined by DNA microarrays. To solve this problem, antibody microarrays are envisioned to replace DNA microarrays in proteome research. These arrays consist of a multitude of different antibodies that are immobilized on a solid support and allow characterization of the protein repertoire of a given sample. However, the production of such antibody microarrays and its application require the provision of highly specific and stable antibodies, possessing high affinity and showing no cross-reactivity. Protein and antibody microarrays can be made to encompass as many as 10,000 samples on a chip within the dimensions of a microscope slide (26).

Monoclonal Antibodies in the Food Industry

Monoclonal antibodies are being used in the wine industry. Odors sometimes observed in spoiled food or corked wines are often the result of microbes present in the wood packaging materials. However, this phenomenon has also been observed in bottled water, suggesting that there may be secondary contaminants, such as residues of pesticides that can affect the quality of any packaged food or beverage. To further the quality assurance of products in the wine industry, a project is being carried out to raise antibodies against a TCA molecule (2,4,6-trichcholoanisole) that is thought to be present in cork stoppers and is responsible for the musty taste in wine. The ELISA assay will be employed to detect trace amounts of the contaminating molecule. Also an immunosensor will be used to electrochemically detect the antibody levels present (27). Monoclonal antibodies have also been developed against the vegetative cells and spores of *Bacillus cereus* (28). This bacterium seems to be implicated in food poisoning and is also responsible for food spoilage. It is impossible for the food industry to exclude *B. cereus* from its products because *B. cereus* cells can survive heat processing and can grow in foods kept at refrigerated storage conditions. Two different antibodies were developed. One was used as a specific capture antibody to destroy the bacterium; the other as a detector antibody that would simply identify the presence of *B. cereus*. The ELISA assay was used to detect and quantify the vegetative cells of this pathogenic organism.

Potato cyst nematodes are pests that destroy the potato food crops. Monoclonal antibodies are being used to assist in the development of the plant's resistance to the nematode (29). Recombinant plant monoclonal antibodies have been engineered to protect poultry against coccidosis infections (30).

Monoclonal Antibodies and Bioterrorism

The same plant biotechnology described above is being developed to create strategic reserves of vaccines and antibodies for infectious agents that could be used in biowarfare. Multiple genes can be engineered in plants intended to provide prolonged immunity against new strains of pathogens that have different mechanisms of action. With this technology, every plant cell will produce the signature protein of a particular biowarfare agent. That protein, in turn, will trigger an immune response in a person who consumes the plant material in an unprocessed or lightly processed form, but it will not cause the disease. These antibodies can prevent infection on surface areas, including nasal passages; clear infectious organisms from the body; identify foreign organisms for destruction; and neutralize and remove toxins. Among the disease-causing substances are several potential bioterrorism agents, such as the botulism toxin, anthrax, Ebola virus, plague, and ricin, a poisonous protein found in the seeds of the castor oil plant. Vaccines for anthrax (*Bacillus anthracis*) and bubonic and pneumonic plague (*Yersinia pestis*), two potentially deadly diseases that can be delivered as airborne agents, are being developed. Preliminary data predicts success in using these plant-derived vaccines (31).

CONCLUDING REMARKS

Antibodies, monoclonal antibodies and antibody derivatives constitute ~20 % of biopharmaceutical products currently in development. Antibodies represent an important and growing class of biotherapeutics. Progress in antibody engineering has allowed the manipulation of the basic antibody structure into its minimal essential functions, and multiple methodologies have emerged for raising and tailoring specificity and functionality. The myriad of monoclonal antibody structures that can be designed and obtained in different formats from various production systems (bacterial, mammalian, and plants) represents a challenge for the recovery and purification of novel immunotherapeutics (3). However, the general use in clinical practice of antibody therapeutics is dependent not only on the availability of products with required efficacy but also on the costs of therapy. As a rule, a significant percentage (50–80%) of the total manufacturing cost of a therapeutic antibody is incurred during downstream processing. The critical challenges posed by the production of novel antibody therapeutics include improving process economics and efficiency to reduce costs, and fulfilling increasingly demanding quality criteria for FDA approval.

BIBLIOGRAPHY

1. Janeway CA JR, Travers P, Walport M, Shlomchik M. Immunobiology. The Immune System in Health and Disease. 6th ed. Churchill Livingstone.
2. Kohler G, Milstein C. Continuous cultures of fused cells secreting antibody of predefined specificity. Nature (London) 1975;256:495–497.
3. Roque AC, Lowe CR, Taipa MA. Antibodies and genetically engineered related molecules: Production and purification. Biotechnol Prog 2004;20:639–654.
4. Kretzschmar T, von Ruden T. Antibody discovery: Phage display. Curr Opin Biotechnol 2002;13:598–602.
5. Fischer R, Twyman RM, Schillberg S. Production of antibodies in plants and their use for global health. Vaccine 2003; 21:820–825.
6. Chen HR. et al. Humanized anti-CD25 monoclonal antibody for prophylaxis of graft-vs-host disease (GVHD) in haploidentical bone marrow transplantation without ex vivo T-cell depletion. Exp. Hematol 2003;31:1019–1025.
7. Hartmann TB, Bazhin AV, Schadendorf D, Eichmuller SB. SEREX identification of new tumor antigens linked to melanoma-associated retinopathy. Int. J. Cancer 2005;10;114:88–93. http://www.licr.org/SEREX.html.
8. Davis DW. et al. Regional effects of an antivascular endothelial growth factor receptor monoclonal antibody on receptor phosphorylation and apoptosis in human 253J B-V bladder cancer xenografts. Cancer Res 2004;64:4601–4610.
9. Christiansen J, Rajasekaran AK. Biological impediments to monoclonal antibody–based cancer immunotherapy. Mol Cancer Ther 2004;3:1493–1501.
10. Koral KF. Update on hybrid conjugate-view SPECT tumor dosimetry and response in 131I-tositumomab therapy of previously untreated lymphoma patients. J Nucl Med 2003;44: 457–464.
11. Villa AM, Berman B. Immunomodulators for skin cancer. J Drugs Dermatol 2004;3:533–539.
12. Von Mehren M, Adams GP, Weiner LM. Monoclonal antibody therapy for cancer. Annu Rev Med 2003;54:343–369.
13. Leget GA, Czuczman MS. Use of rituximAb, the new FDA-approved antibody. Curr Opin Oncol 1998;10:548–551.
14. McLaughlin P. et al. RituximAb chimeric anti-CD20 monoclonal antibody therapy for relapsed indolent lymphoma: Half of patients respond to a four-dose treatment program. J Clin Oncol 1998;16:2825–2833.
15. Leyland-Jones B. TrastuzumAb: hopes and realities. Lancet Oncol 2002;3:137–144.
16. Riethmuller G. et al. Monoclonal antibody therapy for resected Dukes' C colorectal cancer: Seven-year outcome of a multicenter randomized trial. J Clin Oncol 1998;16:1788–1794.
17. Mellstedt H, et al. The therapeutic use of monoclonal antibodies in colorectal carcinoma. Semin Oncol 1991;18:462–477.
18. Khawli LA, Miller GK, Epstein AL. Effect of seven new vasoactive immunoconjugates on the enhancement of monoclonal antibody uptake in tumors. Cancer 1994;73:824–831.
19. Zeidler R, et al. The Fc-region of a new class of intact bispecific antibody mediates activation of accessory cells and NK cells and induces direct phagocytosis of tumour cells. Br J Cancer 2000;83:261–266.
20. Trauth BC, et al. Monoclonal antibody-mediated tumor regression by induction of apoptosis. Science 1989;245:301–305.
21. Yang XD, et al. Eradication of established tumors by a fully human monoclonal antibody to the epidermal growth factor receptor without concomitant chemotherapy. Cancer Res 1999;59:1236–1243.
22. Agus DB, et al. Targeting ligand-activated ErbB2 signaling inhibits breast and prostate tumor growth. Cancer Cell 2002;2:127–137.
23. Estey EH, et al. Experience with gemtuzumAb ozogamycin ("mylotarg") and all-*trans* retinoic acid in untreated acute promyelocytic leukemia. Blood 2002;99:4222–4224.
24. Bander NH, et al. Targeted systemic therapy of prostate cancer with a monoclonal antibody to prostate-specific membrane antigen. Semin Oncol 2003;30:667–676.
25. Witzig TE, et al. Randomized controlled trial of yttrium-90–labeled ibritumomAb tiuxetan radioimmunotherapy versus rituximAb immunotherapy for patients with relapsed or refractory low-grade, follicular, or transformed B-cell non-Hodgkin's lymphoma. J Clin Oncol 2002;20:2453–2463.
26. Angenendt P, et al. Seeing better through a MIST: evaluation of monoclonal recombinant antibody fragments on microarrays. Anal Chem 2004;76:2916–2921.
27. Sanvicens N, Varela B, Marco MP. Determination of 2,4,6-trichloroanisole as the responsible agent for the musty odor in foods. 2. Immunoassay evaluation. J Agric Food Chem 2003;51: 3932–3939.
28. Charni N, et al. Production and characterization of monoclonal antibodies against vegetative cells of Bacillus cereus. Appl Environ Microbiol 2000;66:2278–2281.
29. Fioretti L, et al. Monoclonal antibodies reactive with secreted-excreted products from the amphids and the cuticle surface of Globodera pallida affect nematode movement and delay invasion of potato roots. Int J Parasitol 2002;32:1709–1718.
30. Refega S, et al. Production of a functional chicken single-chain variable fragment antibody derived from caecal tonsils B lymphocytes against macrogamonts of Eimeria tenella. Vet Immunol Immunopathol 2004;97:219–230.
31. Petrenko VA, Sorokulova IB. Detection of biological threats. A challenge for directed molecular evolution. J Microbiol Methods 2004;58:147–168.

See also BORON NEUTRON CAPTURE THERAPY; IMMUNOTHERAPY.

MOSFET. See ION-SENSITIVE FIELD-EFFECT TRANSISTORS.

MRI. See MAGNETIC RESONANCE IMAGING.

MUSCLE ELECTRICAL ACTIVITY. See ELECTROMYOGRAPHY.

MUSCLE TESTING, REHABILITATION AND. See REHABILITATION AND MUSCLE TESTING.

MUSCULOSKELETAL DISABILITIES. See REHABILITATION, ORTHOTICS FOR.